Dub-S2ST: Textless Speech-to-Speech Translation for Seamless Dubbing

Jeongsoo Choi* **Jaehun Kim*** **Joon Son Chung**
Korea Advanced Institute of Science and Technology
{jeongsoo.choi, kjaehun, joonson}@kaist.ac.kr

Abstract

This paper introduces a cross-lingual dubbing system that translates speech from one language to another while preserving key characteristics such as duration, speaker identity, and speaking speed. Despite the strong translation quality of existing speech translation approaches, they often overlook the transfer of speech patterns, leading to mismatches with source speech and limiting their suitability for dubbing applications. To address this, we propose a discrete diffusion-based speech-to-unit translation model with explicit duration control, enabling time-aligned translation. We then synthesize speech based on the translated units and source speaker's identity using a conditional flow matching model. Additionally, we introduce a unit-based speed adaptation mechanism that guides the translation model to produce speech at a rate consistent with the source, without relying on any text. Extensive experiments demonstrate that our framework generates natural and fluent translations that align with the original speech's duration and speaking pace, while achieving competitive translation performance.

1 Introduction

Recent advancements in translation systems and speech technologies have enabled a vast amount of multimedia content to support multiple languages through automated speech dubbing. Cross-lingual dubbing (Federico et al., 2020a; Wu et al., 2023), which replaces speech audio of one language with that of another, allows global audiences to consume content in their native languages. While this significantly reduces language barriers, ensuring effective dubbing requires meeting several specific criteria: maintaining duration, speaker identity, and speaking speed (Brannon et al., 2023).

Traditional dubbing systems typically employ a cascade of Automated Speech Recognition (ASR) (Amodei et al., 2016; Baevski et al., 2020; Radford et al., 2023), Neural Machine Translation (NMT) (Johnson et al., 2017; Stahlberg, 2020; Fan et al., 2021; Costa-jussà et al., 2022), and Text-to-Speech (TTS) (Wang et al., 2017; Ren et al., 2021; Wang et al., 2023a; Tan et al., 2024b) modules. Although these cascaded systems demonstrate promising translation quality, they inherently lose critical speech-related information, such as speaker identity and prosody, due to the intermediate text representations (Swiatkowski et al., 2023). Moreover, because text lacks precise duration information, these systems struggle to accurately match the duration and speaking pace of the original speech (Sahipjohn et al., 2024). As a result, even with post-processing, the output often remains misaligned or unnatural, limiting their effectiveness for real-world dubbing (Effendi et al., 2022).

To address these limitations, Speech-to-Speech Translation (S2ST) systems (Jia et al., 2019, 2022a; Barrault et al., 2025) have emerged and evolved into textless S2ST (Lee et al., 2021; Li et al., 2023; Kim et al., 2024), aiming to translate speech directly without intermediate text. While recent approaches have achieved translation quality comparable to cascaded systems, most of them lack the capability to control speech duration during translation. As a result, the output typically requires post-processing like manual stretching or contracting to match the original speech duration for dubbing purposes. However, this process can degrade speech quality and lead to unnatural prosody and speaking pace.

A key challenge underlying this issue lies in the limitations of existing training datasets. High-quality dubbing demands not only accurate translation but also faithful preservation of the source speech's voice characteristics and speaking speed. However, it is inherently difficult to construct large-scale datasets containing the same speaker uttering aligned content in multiple languages. Conse-

*Equal contribution.

quently, most S2ST datasets rely on synthesized target speech (Jia et al., 2022b) or web-crawled data (Duquenne et al., 2023a; Barrault et al., 2025), which prioritize linguistic fidelity over speech-related information consistency. These datasets tend to exhibit discrepancies in speaker identity and speaking speed, making it challenging for models to learn how to preserve such attributes. Despite progress in translation accuracy, existing S2ST models trained on these datasets remain suboptimal for seamless dubbing.

In response, we propose Dub-S2ST, a novel textless S2ST framework specifically designed for dubbing applications, effectively leveraging existing datasets. To mitigate the ambiguity caused by discrepancies between source and target speech, we eliminate variations in the target that deviate from the paired source. Specifically, we first convert continuous speech to discrete units that retain rich semantic features and minimal acoustic variations (Lakhotia et al., 2021; Lee et al., 2021). We then apply our unit-based speed adaptation strategy to adjust the target's speaking rate to the source. Using the processed data, we develop a speech-to-unit translation model trained with a discrete diffusion objective (Austin et al., 2021). We also incorporate Diffusion Transformer (Peebles and Xie, 2023), which allows the model to accurately predict speech units conditioned on diffusion timestep. Moreover, our model inherently supports duration control by using predetermined lengths based on the duration of source speech. Finally, we incorporate a conditional flow matching (CFM) (Lipman et al., 2023; Mehta et al., 2024)-based synthesizer that generates high-quality speech conditioned on the translated speech units and the original source speech, closely resembling the original speaker's identity.

Through extensive evaluation, our proposed framework demonstrates superior preservation of duration, speaker identity, and speaking speed, while maintaining competitive translation accuracy. Ablation studies further validate the effectiveness of each component in improving dubbing quality. To the best of our knowledge, Dub-S2ST is the first textless S2ST framework tailored for seamless automatic dubbing that preserves both speaker identity and speaking speed.

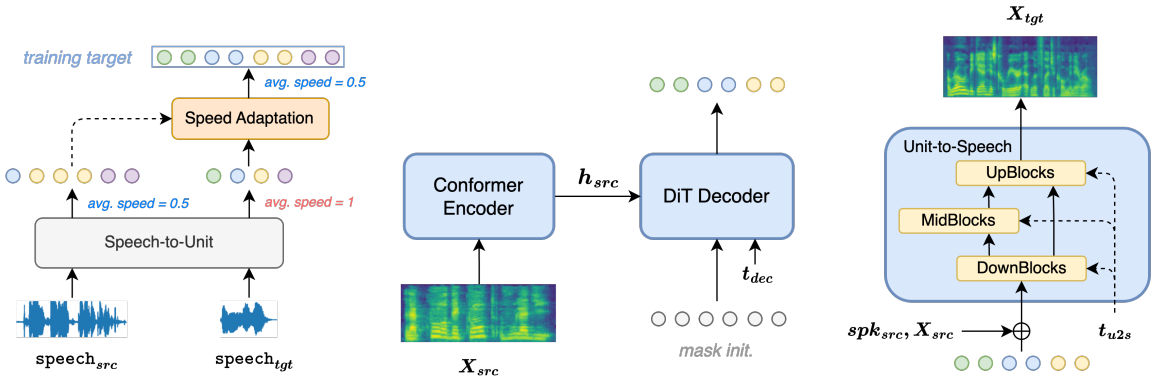
2 Related Works

2.1 Cross-lingual Dubbing

Dubbing is a post-production process in which the original spoken dialogue in multimedia content is replaced with speech in another language, while preserving the temporal alignment and naturalness of the original speech (Orero, 2004). Early automatic dubbing systems typically adopt cascaded S2ST architecture, combining ASR, NMT, and TTS (Dureja and Gautam, 2015). While maintaining this cascaded pipeline, recent efforts have focused on enhancing each component. Some approaches focus on improving prosodic alignment, aiming to synchronize the prosody of the generated speech with the original (Federico et al., 2020b). Others leverage existing TTS models by modifying the duration module to generate speech that matches the original duration (Effendi et al., 2022). Despite these advancements, relying on text as intermediate representation inherently limits temporal flexibility, highlighting the need for textless approaches that better preserve the naturalness of the source speech.

2.2 Speech-to-Speech Translation (S2ST)

Speech-to-Speech Translation (S2ST) aims to convert source speech into a target language while preserving linguistic content. Early systems adopted a cascaded architecture, integrating ASR, MT, and TTS modules (Federico et al., 2020b; Lakew et al., 2022). To mitigate error propagation and latency issues inherent in cascaded systems, direct S2ST approaches have been introduced (Jia et al., 2019, 2022a), demonstrating the feasibility of end-to-end speech translation. To further eliminate reliance on intermediate text, textless S2ST methods have emerged. S2UT (Lee et al., 2021) proposes an autoregressive (AR) translation model that predicts deduplicated discrete speech units (Lakhotia et al., 2021), which are then used to synthesize target speech. UTUT (Kim et al., 2024) extends this framework to support many-to-many language translation. Subsequent works like TranSpeech (Huang et al., 2023) explore non-autoregressive (NAR) models to achieve faster decoding compared to AR models, and DiffNorm (Tan et al., 2024a) enhances TranSpeech by introducing techniques to normalize acoustic variations in the target speech. More recently, CTC-S2UT (Fang et al., 2024) incorporates CTC (Graves et al., 2006)-based unit reduction to improve trans-



(a) Data preparation (b) Speech-to-Unit translation (c) Unit-to-Speech synthesis

Figure 1: Dub-S2ST framework. (a) *avg. speed* indicates average unit speed calculated in unit-based speed adaptation. (b) h_{src} indicates the encoded source speech features from the encoder, and t_{dec} is the timestep information used to train discrete-diffusion decoder. (c) spk_{src}, X_{src} represent speaker embedding and mel-spectrogram from the source speech as conditions, respectively. t_{u2s} is the timestep information used to train unit-to-speech synthesizer.

lation performance. Unlike previous approaches, we leverage a NAR-based model without deduplicating speech units. This design preserves temporal information and enables explicit control over output duration, making our framework naturally suitable for dubbing scenarios.

2.3 S2ST Datasets

Recent advancements in S2ST have been facilitated by the development of specialized datasets. VoxPopuli (Wang et al., 2021) is a multilingual corpus containing aligned speech pairs derived from simultaneous interpretation by human interpreters at European Parliament events. This dataset provides realistic translation pairs, however, it is limited in terms of domain coverage and scalability. A common approach to handle these issues is synthesizing target speech from translated text (Jia et al., 2019, 2022a). For instance, CVSS (Jia et al., 2022b) is built by converting the text from the speech-to-text translation corpus CoVoST2 (Jia et al., 2022b) into speech using a pretrained TTS model. Another approach is based on data mining. SeamlessAlign (Barrault et al., 2025) leverages web-crawled multimodal translation data and employs a unified speech-text similarity model (Duquenne et al., 2023b) to effectively pair speech segments, resulting in approximately 29,000 hours of S2ST data. However, these datasets still lack consistency in speech attributes between source and target. Our goal is to develop a framework that can generate dubbed speech that closely follows the source, even under such conditions.

3 Method

In this section, we explain the data preparation, model architecture and training objective, and final synthesis of the translated speech. The overall architecture of our model is illustrated in Fig 1.

3.1 Data Preparation

Speech Unit Extraction. The choice of target speech representation plays a critical role in determining the quality and accuracy of S2ST. While continuous features allow for straightforward generation, they are often susceptible to noise and mispronunciations, resulting in reduced intelligibility and naturalness. In contrast, discrete features leverage pretrained speech Self-Supervised Learning (SSL) models followed by quantization (Hsu et al., 2021; Lee et al., 2021). Empowered by large-scale SSL training and discretization, this approach effectively captures phonetic information while minimizing speaker-dependent attributes such as timbre and pitch (Lakhotia et al., 2021).

To focus explicitly on the linguistic components of speech during translation, the proposed method adopts units extracted using mHuBERT (Lee et al., 2022), followed by k-means clustering for quantization. This cascaded unit extraction process facilitates robust representation of linguistic content while suppressing paralinguistic variations.

Unit-Based Speed Adaptation. Speaking speed is an essential factor in dubbing, as perceptual quality significantly degrades when the speed of translated speech deviates from the original (Orero, 2004). While signal processing techniques can be applied

to adjust speech duration, they often compromise naturalness and intelligibility. Another strategy is to guide models using syllable- or phoneme-based speed metrics, since these provide an indication of speaking pace across languages (Barrault et al., 2025); however, such methods are inapplicable in textless systems.

To address these limitations, we propose a unit-based speed adaptation method that adjusts the repetition of speech units based on the speaking speed ratio between source and target speech. This method is inspired by unit deduplication (*i.e.*, repetition removal) used in recent S2ST models (Lee et al., 2021; Huang et al., 2023; Fang et al., 2024) to preserve synthesis quality while minimizing redundancy. We hypothesize that the reduced sequence $\hat{L}$ captures a distinct set of pronunciations, and the ratio against original length L serves as an implicit estimate of speaking speed $r = \frac{\hat{L}}{L}$.

After extracting speech unit sequences from both source and target, we adjust the target sequence by applying the speed ratio $\frac{r_{src}}{r_{tgt}}$, modifying the number of unit repetitions to align the target speed with the source. The speed is normalized by the average speed of each language to mitigate cross-linguistic differences. This adaptation preserves linguistic content while controlling speaking rate. Note that, as depicted in Figure 1, the speed adaptation method does not force the target speech units to have the same length as the source. It rather changes the rate of repeating units, so that it only alters the implicit speaking pace. Training the model on speed-aligned sequences enables it to generate translated speech that naturally and consistently mirrors the source speaking speed.

3.2 Speech-to-Unit Translation

Accurate dubbing requires the translated speech to match the duration of the source utterance. To fulfill this requirement, we propose a speech-to-unit translation model whose decoder is implemented as a discrete diffusion (Austin et al., 2021) generator conditioned on the source speech. The decoder leverages Diffusion Transformer (DiT) (Peebles and Xie, 2023) layers that operate on variable-length sequences, with additional cross-attention to integrate source speech features throughout the generation. During inference, the decoder takes a fully masked unit sequence with length identical to that of the source speech as its initial input and iteratively transforms it into speech units.

The training is conducted by first masking units based on a mask schedule $\gamma(t_{dec})$, where both t_{dec} and $\gamma(t_{dec})$ range from $(0, 1)$. Each unit is independently masked with probability $\gamma(t_{dec})$ and remains unmasked with probability $1 - \gamma(t_{dec})$, and the decoder predicts the original target units from the partially masked sequence. The decoder is optimized using the following cross-entropy loss:

$$\mathcal{L} = -\mathbb{E}\left[\frac{1}{|M|}\sum_{i\in M}\log p_\theta(x_0^i \mid x_t, h_{\text{src}}, t_{dec})\right] \tag{1}$$

Here, M is the set of indices corresponding to masked units, x_0^i is the ground-truth speech unit, x_t is the partially masked input sequence at timestep t_{dec}, and h_{src} is the encoded representation from the Conformer encoder. During training, the decoder predicts speech units from partially masked target sequence. During inference, it receives a fully masked sequence whose length matches that of the source speech units, thereby generating translated speech aligned in length with the source. The loss is applied only to the masked units, and we further investigate the impact of this in Section 5.

To enhance the decoder's reliance on source speech representations, we initialize the encoder from a pretrained autoregressive speech-to-unit translation model (Lee et al., 2022) and fine-tune the entire weights. Additionally, we apply label-smoothing (Szegedy et al., 2016) with a factor empirically set to 0.01 to improve generalization.

3.3 Unit-to-Speech Synthesis

A key factor in the perceptual quality of dubbing is the similarity between the translated speech and the original utterance. To convert semantic units to speech while preserving the acoustic characteristics of the source speech, we employ a unit-to-speech synthesizer based on Optimal Transport Conditional Flow Matching (OT-CFM). The synthesizer is implemented as a U-Net (Rombach et al., 2022) architecture, where each layer is a block comprising Convolutional and Transformer layers. Downsampling and upsampling are performed in the latent space, allowing the model to efficiently reconstruct fine-grained temporal structure.

We train the model using the OT-CFM objective, which defines a time-dependent vector field that transports a sample from a simple prior distribution $x_0 \sim \mathcal{N}(0, I)$ to a data sample $x_1 \sim q$ via linear displacement interpolation:

$$\varphi_t = (1 - (1 - \sigma_{\min})t_{u2s})\, x_0 + t_{u2s} x_1 \tag{2}$$

The decoder is trained to minimize the difference between the predicted and target velocities:

$$\mathcal{L} = \mathbb{E}_{t_{u2s}, x_0, x_1} \|u_t(\varphi_t \mid x_1) - v_\theta(\varphi_t | t_{u2s}, \boldsymbol{c})\|^2 \quad (3)$$

where the ground-truth velocity is given by:

$$u_t^{\text{OT}}(\varphi_t^{\text{OT}} \mid x_1) = x_1 - (1 - \sigma_{\min})x_0 \quad (4)$$

The condition $\boldsymbol{c}$ consists of the unit embedding sequence from source and target speech, the source speaker embedding spk_{src} extracted with a pretrained speaker verification model (Wang et al., 2023b), and the source mel-spectrogram X_{src}. The mel-spectrogram is padded to match the length of the unit sequence, while the speaker embedding is repeated accordingly. These features are concatenated channel-wise to enable in-context learning: the sampled prior is transformed into a mel-spectrogram conditioned on both speaker identity and prosodic information from the source speech.

Although directly training the module on S2ST data is possible, such datasets often contain noise and reverberation, which can impair synthesis quality. To address this, we initialize the model from a TTS model (Du et al., 2024) trained on multilingual corpus, and fine-tune it with necessary adaptations for our semantic unit input. This approach allows for robust zero-shot synthesis across diverse speakers and languages. The generated mel-spectrogram is then converted to audible waveform via a pretrained HiFi-GAN (Kong et al., 2020).

4 Experiment

4.1 Dataset

CVSS-C (Jia et al., 2022b) is a widely-used dataset for speech-to-speech translation, consisting of 21 languages to English translations where English speech is generated by a single-speaker TTS model. The proposed method is trained and evaluated with French-English (fr-en) subset, due to the abundance of samples compared to other language pairs. The fr-en subset contains 207,364 (train) / 14,759 (dev) / 14,759 (test) pairs of source and target speech samples, totaling 264 hours. The experiment utilizes train and dev split for training and validation, and test split for final evaluation.

4.2 Implementation Details

Preprocessing. Audio samples are resampled to 16kHz and preprocessed with Voice Activity Detection (VAD) tool[1] to remove unnecessary silence and paralinguistic information at the beginning and end. The samples are then processed with pretrained mHuBERT (Lee et al., 2021) and k-means clustering model to obtain discrete units[2].

Architecture. To maintain consistency with prior works on speech-to-unit translation models, we design our model with 12 Conformer encoder layers and 6 DiT layers, totaling 61M parameters. The unit-to-speech synthesizer consists of 4 Down, Mid, and UpBlock, where each is a cascade of 1D Convolution and Transformer layer.

Training. The speech-to-unit translation model is trained with a total batch size of 3,200 seconds for 100k updates. We use 8 RTX A5000 GPUs for training, and the total training takes approximately 10 hours. We optimize the model with the AdamW (Loshchilov and Hutter, 2019) optimizer and applied a dropout rate of 0.3. The learning rate is warmed up for the first 10k steps to a peak of 1×10^{-3}, and then decayed using an inverse square root schedule. We implement our approach using Fairseq (Ott et al., 2019). The unit-to-speech module is initialized from CosyVoice-300M (Du et al., 2024) and fine-tuned with fixed learning rate of 1×10^{-4} for 200k steps using LRS3 (Afouras et al., 2018) dataset, an English multi-speaker corpus.

4.3 Evaluation Metrics

ASR-BLEU (Lee et al., 2022) is a widely adopted metric for assessing S2ST quality. It measures translation quality by transcribing the generated speech using a pretrained ASR model (Baevski et al., 2020) and comparing it with the ground-truth text to compute BLEU (Post, 2018) [3].

BLASER 2.0 (Dale and Costa-jussà, 2024) serves as an automatic measure for assessing semantic similarity between source and generated speech. We adopt its reference-free variant, BLASER 2.0-QE[4], which estimates translation quality without reference text unlike BLEU score.

SIM evaluates the speaker similarity between the generated and original speech. We use a pretrained speaker verification model (Chen et al., 2022b), based on WavLM-Large (Chen et al., 2022a), to extract speaker embedding vectors and calculate cosine similarity between the two.

[1] https://github.com/snakers4/silero-vad

[2] https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_speech/docs/textless_s2st_real_data.md

[3] https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_to_speech/asr_bleu

[4] https://huggingface.co/facebook/blaser-2.0-qe

Duration Controllable	Method	ASR-BLEU ↑	BLASER 2.0 ↑	SIM ↑	DNSMOS ↑
✗	S2UT (Lee et al., 2022)	24.54	3.784	0.036	3.922
	CTC-S2UT (Fang et al., 2024)	24.51	3.785	0.037	3.908
	UTUT (Kim et al., 2024)†	26.49	3.840	0.036	3.927
	w/ Zero-shot Vocoder (Choi et al., 2024)†	26.33	3.882	0.145	3.101
✓	TranSpeech (Huang et al., 2023)‡	18.03	-	-	-
	DiffNorm (Tan et al., 2024a)‡	19.53	-	-	-
	Dub-S2ST-single (Ours)	23.88	3.813	0.036	**3.945**
	w/o speed adaptation	22.10	3.766	0.035	3.909
	Dub-S2ST (Ours)	**24.16**	**3.839**	**0.266**	3.693

Table 1: Performance comparisons with state-of-the-art textless S2ST methods on CVSS-C dataset. †Multilingual translation model trained with a larger model size and dataset. ‡The scores are reported from the original papers.

Method	DC@0.2	DC@0.4	SC@0.2	SC@0.4	S. Corr
S2UT (Lee et al., 2022)	64.53	93.65	62.01	90.02	0.222
w/o unit deduplication	56.45	89.96	57.88	85.68	0.254
w/ pos. emb. (Wu et al., 2023)	79.67	99.20	61.76	88.71	0.355
w/ pos. emb. (Le et al., 2024)	80.78	99.16	61.46	88.61	0.367
CTC-S2UT (Fang et al., 2024)	65.33	94.34	62.70	90.38	0.251
Dub-S2ST-single (Ours)	**100.00**	**100.00**	**71.93**	**96.77**	**0.614**
w/o speed adaptation	**100.00**	**100.00**	66.65	92.16	0.388

Table 2: The performance comparisons about speech duration and speed. DC@p and SC@p indicate duration and speed compliance with range p, while S. Corr denotes correlation between the syllable speed of source and generated speech.

Method	Naturalness	Translation Consistency	Speed Consistency
S2UT (Lee et al., 2022)	2.60 ± 0.18	3.32 ± 0.16	3.62 ± 0.19
CTC-S2UT (Fang et al., 2024)	2.52 ± 0.19	3.56 ± 0.15	3.55 ± 0.16
Dub-S2ST-single (Ours)	**3.37** ± 0.17	**3.80** ± 0.13	**3.98** ± 0.16

Table 3: MOS evaluation.

DNSMOS (Deep Noise Suppression Mean Opinion Score) (Reddy et al., 2021) is an automated perceptual speech quality assessment of generated speech[5]. The quality is estimated with a score in the range of [1, 5] where the larger value indicates higher speech quality.

Duration Compliance (DC) (Wu et al., 2023) calculates the portion of generated speech whose duration ratio with the source lies within certain range. This indicates how S2ST system preserves the duration when generating translated speech.

Speed Compliance (SC) captures a similar aspect to DC but is based on the ratio between the speed of speech, measured in syllables per second[6]. This metric reflects how closely the speed of the generated speech aligns with that of the source.

5 Experimental Results

5.1 Quantitative Comparison

Translation Quality. Table 1 presents various evaluation results comparing the proposed method against existing baselines. Dub-S2ST-single, which employs the unit vocoder of S2UT for fair comparison with single-speaker approaches, outperforms all existing duration-controllable methods, achieving a BLEU score of 23.88. Disabling our proposed speed adaptation strategy leads to performance drops across all metrics, highlighting its effectiveness. A more detailed analysis of the speed adaptation is presented in the following section.

In addition, the last row in the table reports the performance of Dub-S2ST using our proposed multi-speaker unit-to-speech synthesizer. It shows superior speaker identity preservation, outperforming all baselines in terms of speaker similarity. Furthermore, Dub-S2ST achieves highly competitive ASR-BLEU and BLASER 2.0 scores compared to S2UT and CTC-S2UT, which lack duration control. This indicates that our model effectively captures the semantic information from the source speech and transfers it to the translated speech, while maintaining the duration and identity.

Duration and Speed Analysis. We evaluate our model's performance in duration control by measuring the generated speech duration relative to the source, as shown in Table 2. In addition to standard S2ST baselines, we compare against methods that incorporate duration control: VideoDubber (Wu et al., 2023), which employs additional positional embeddings, and TransVIP (Le et al., 2024), which introduces isochrony positional embedding based on voice activity information. While existing methods achieve reasonable compliance within a 40%

[5]We use a model trained with ITU-T P.808: https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS

[6]https://github.com/facebookresearch/seamless_communication/blob/main/docs/expressive

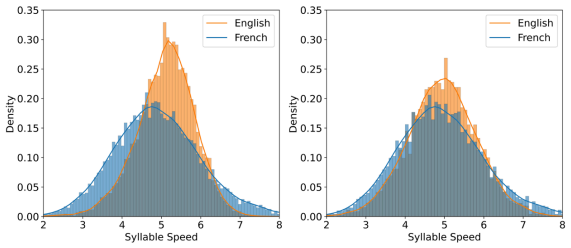


Figure 2: Change in syllable speed distribution on the CVSS-C dataset (a) before and (b) after applying our speed adaptation.

vs Source syllable speed	Speed adapt.	Corr
Source unit speed		0.606
Target syllable speed	✗	0.235
Target syllable speed	✓	**0.519**

Table 4: Effectiveness of our speed adaptation strategy.

threshold, they struggle under a stricter 20% constraint. In contrast, our model achieves exact duration matching through explicit length initialization in the decoder, yielding 100% compliance.

The proposed model's advantage becomes more evident in the speed analysis, where our model, both with and without speed adaptation, outperforms all baselines in speed compliance. This indicates that, when conditioned on the source duration, our model not only matches the duration but also the speaking pace of the source. The proposed speed adaptation further enhances this alignment. Overall, these results underscore the robustness of our approach in controlling speech duration and speed, both of which serve as critical factors in dubbing.

5.2 Human Evaluation

We conducted a Mean Opinion Score (MOS) survey to assess the perceptual quality of the generated speech, as presented in Table 3. 15 professional listeners rated samples on a scale from 1 (poor) to 5 (excellent). To ensure fair comparison, samples with durations differing from the source speech were manually adjusted to match the original.

The speed consistency evaluation confirms that our proposed speed adaptation method effectively aligns the speaking speed between source and target speech, producing translated speech at a pace closely matching the source. This implicit synchronization contributes to improving the translation consistency, since it helps the model generate sequences at a similar speed to the source when given

Ground Truth:	none of this happened
80% duration:	nothing took place
100% duration:	nothing of that took place
120% duration:	nothing of all that took place
Ground Truth:	most authors emphasize the extravagance of certain plans
80% duration:	most of the authors insist on the extravagance of some plans
100% duration:	most of the authors insist on the extravagance of some plans
120% duration:	most of the authors insist on the extravagancy of certain plans

Figure 3: Examples of ASR transcribed translation outputs of our model with varying target durations.

Duration Ratio	0.8	0.9	1.0	1.1	1.2
Relative # Chars	0.851	0.929	1.000	1.064	1.133

Table 5: Effect of explicit duration control.

its duration, minimizing unnecessary pauses and repetitive words. Notably, significant differences in speech naturalness emerged during the evaluation. Baseline samples required manual waveform adjustments, leading to considerable degradation in perceptual quality. Conversely, our proposed approach yielded accurate, time-aligned speech outputs, eliminating the need for post-processing adjustments and maintaining high naturalness suitable for dubbing applications.

5.3 Discussion

How well does unit speed reflect syllable speed? Syllable rate refers to the number of syllables spoken per second and is commonly used for measuring speaking speed. As a preliminary analysis, we measure the correlation between our unit-based speech rate estimation and syllable-based speed derived from transcripts. As shown in Table 4, the proposed metric strongly correlates with ground-truth syllable speed, validating its reliability.

How effective of unit-based speed adaptation? Figure 2 illustrates the effect of the proposed speed adaptation strategy. In Figure 2 (a), the syllable speed distribution of English utterances deviates from that of French. However, after applying speed adaptation, the English distribution in Figure 2 (b) closely aligns with the French distribution. This indicates that the proposed adaptation method effectively adjusts the speaking pace of the target.

Can explicit duration control truly change the translated content? As illustrated in Figure 3, our model responds to different duration prompts by producing semantically consistent outputs with nuanced variations in structure and phrasing. For example, "none of this happened" is rendered as

Decoder	Schedule	NFE				
		1	4	16	64	256
Transformer	✗	2.61	-	-	-	-
Transformer	Cosine	**13.52**	19.42	20.10	20.23	20.33
Transformer	Linear	11.83	20.28	21.13	21.37	21.36
DiT	Linear	12.09	**20.57**	**21.65**	**22.10**	**22.27**

Table 6: Ablation study on the model architecture and masking schedule of the speech-to-unit translation model, evaluated using ASR-BLEU.

Loss computation	ASR-BLEU	BLASER 2.0
All	21.46	3.742
Masked	**22.10**	**3.766**
Masked (non-trivial)	21.46	3.751

Table 7: Ablation study on loss computation during speech-to-unit translation model training.

"nothing took place" at 80% duration and "nothing of all that took place" at 120%, each expressing the same meaning without repetition or loss of information. Moreover, the relative number of characters of the translated outputs with different durations in Table 5 validates that the model flexibly generates semantics that fit the given duration, rather than forcing a fixed translation to fit varying lengths.

5.4 Ablation Study

Model architecture and Masking schedule. To assess how different model architectures and masking schedules affect the speech-to-unit translation model of Dub-S2ST, we compare several design choices, as shown in Table 6. The last row presents ASR-BLEU performances from Dub-S2ST-single without speed adaptation, evaluated across varying numbers of function evaluations (NFE). Replacing our DiT decoder with a standard Transformer decoder consistently degrades translation quality across all NFEs, indicating that incorporating diffusion timestep into the model benefits translation learning. We also examine the effect of the masking schedule by comparing our linear schedule with a cosine schedule. The results show that the linear schedule yields overall better performance, suggesting its effectiveness without losing translation quality. Based on the latency-performance trade-off, we choose the NFE of 64 for our evaluation.

Loss computation. In Table 7, we examine the impact of different loss computation strategies on our model's performance. Our findings indicate that computing loss on all units (*i.e.*, predicting both masked and non-masked units) results in lower translation performance than masked prediction. This outcome is primarily due to the model's tendency to focus on easier predictions, leading to ineffective training. Calculating loss only on masked units, as implemented in our method, yields the best performance across all translation quality metrics. We also evaluated grouped masking strategy, where loss is not computed on masked units if at least one unit in the repeating group is unmasked, but it results in lower performance.

Model	ASR-BLEU	SIM	DNSMOS
Dub-S2ST-single	23.88	0.036	**3.945**
w/ Zero-shot Vocoder (Choi et al., 2024)	23.76	0.154	3.088
w/ CosyVoice VC (Du et al., 2024)	23.09	**0.315**	3.787
Dub-S2ST	**24.16**	0.266	3.693

Table 8: Ablation study on unit-to-speech module.

Unit-to-speech module. To evaluate our unit-to-speech synthesizer's ability to preserve speaker identity, we compare the performance with a zero-shot vocoder proposed in AV2AV (Choi et al., 2024). Additionally, we applied a cross-lingual voice conversion model (Du et al., 2024) to our single-speaker model and compare the results. As shown in Table 8, our unit-to-speech model achieves the highest translation quality, even outperforming the single-speaker model. While the zero-shot vocoder maintains reasonable translation quality, it shows poor speaker similarity. On the other hand, using a separate voice conversion model shows a significant drop in ASR-BLEU. Based on qualitative analysis, we hypothesize that the process of voice conversion introduces over-smoothing of pronunciation that leads to loss in intelligibility and ultimately translation quality.

6 Conclusion

In this paper, we introduce an S2ST system suitable for dubbing applications. The proposed model generates translated speech with accurate content while preserving the duration and speaking speed of the source. This is achieved by the unique design that can generate speech with arbitrary duration, and speed adaptation that mitigates discrepancies between source and target speech. Extensive experiments with systematic ablations demonstrate that Dub-S2ST outperforms the existing baselines and verify its applicability to cross-lingual dubbing.

7 Limitations

While the proposed model offers an effective approach to cross-lingual dubbing, it is trained on speech recorded in controlled environments. This suggests that training on larger and more diverse datasets may be required for application to in-the-wild scenarios. Moreover, the proposed model is trained on a sentence-level speech corpus, and therefore a segmentation process is required before being utilized in other applications.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT, RS-2025-02215122, Development and Demonstration of Lightweight AI Model for Smart Homes and RS-2022-II220989, Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling).

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proc. ICML*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Proc. NeurIPS*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*.

William Brannon, Yogesh Virkar, and Brian Thompson. 2023. Dubbing in practice: A large scale study of human localization with insights for automatic dubbing. *Trans. of the Association for Computational Linguistics*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022b. Large-scale self-supervised speech representation learning for automatic speaker verification. In *Proc. ICASSP*.

Jeongsoo Choi, Se Jin Park, Minsu Kim, and Yong Man Ro. 2024. Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. In *Proc. CVPR*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

David Dale and Marta Costa-jussà. 2024. Blaser 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of EMNLP*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2023a. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. In *Proc. ACL*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023b. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Mahak Dureja and Sumanlata Gautam. 2015. Speech-to-speech translation: A review. *International Journal of Computer Applications*.

Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico. 2022. Duration modeling of neural tts for automatic dubbing. In *Proc. ICASSP*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*.

Qingkai Fang, Zhengrui Ma, Yan Zhou, Min Zhang, and Yang Feng. 2024. Ctc-based non-autoregressive textless speech-to-speech translation. In *Findings of ACL*.

Marcello Federico, Robert Enyedi Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020a. From speech-to-speech translation to automatic dubbing. *IWSLT 2020*.

Marcello Federico, Yogesh Virkar, Robert Enyedi, and Roberto Barra-Chicote. 2020b. Evaluating and optimizing prosodic alignment for automatic dubbing. In *Proc. Interspeech*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

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

Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, and Zhou Zhao. 2023. Transpeech: Speech-to-speech translation with bilateral perturbation. In *Proc. ICLR*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *Proc. ICML*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. Cvss corpus and massively multilingual speech-to-speech translation. In *Proc. LREC*.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *Proc. Interspeech*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. of the Association for Computational Linguistics*.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2024. Textless unit-to-unit training for many-to-many multilingual speech-to-speech translation. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*.

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

Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isometric mt: Neural machine translation for automatic dubbing. In *Proc. ICASSP*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Trans. of the Association for Computational Linguistics*.

Chenyang Le, Yao Qian, Dongmei Wang, Long Zhou, Shujie Liu, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Sheng Zhao, et al. 2024. Transvip: Speech to speech translation system with voice and isochrony preservation. In *Proc. NeurIPS*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proc. ACL*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. 2021. Textless speech-to-speech translation on real data. In *Proc. NAACL*.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. 2023. Textless direct speech-to-speech translation with discrete speech representation. In *Proc. ICASSP*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *Proc. ICLR*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-tts: A fast tts architecture with conditional flow matching. In *Proc. ICASSP*.

Pilar Orero. 2004. *Topics in audiovisual translation*. John Benjamins Publishing Company.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proc. CVPR*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. ICASSP*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*.

Neha Sahipjohn, Ashishkumar Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Rajiv Ratn Shah. 2024. Dubwise: Video-guided speech duration control in multimodal llm-based text-to-speech for dubbing. In *Proc. Interspeech*.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*.

Jakub Swiatkowski, Duo Wang, Mikolaj Babianski, Patrick Lumban Tobing, Ravichander Vipperla, and Vincent Pollet. 2023. Cross-lingual prosody transfer for expressive machine dubbing. In *Proc. Interspeech*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. CVPR*.

Weiting Tan, Jingyu Zhang, Lingfeng Shen, Daniel Khashabi, and Philipp Koehn. 2024a. Diffnorm: Self-supervised normalization for non-autoregressive speech-to-speech translation. In *Proc. NeurIPS*.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024b. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. ACL*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. Cam++: A fast and efficient network for speaker verification using context-aware masking. In *Proc. Interspeech*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. 2023. Videodubber: Machine translation with speech-aware length control for video dubbing. In *Proc. AAAI*.