2025년도 한국통신학회 추계종합학술발표회

더빙을 위한 음성 직접 번역에 관한 연구

김재훈, 정준선*
한국과학기술원, *한국과학기술원

kjaehun@kaist.ac.kr, *joonson@kaist.ac.kr

A Study on textless speech-to-speech translation systems

Jaehun Kim, Joon Son Chung*
KAIST., *KAIST.

Abstract

This paper introduces a multi-lingual textless speech-to-speech translation system for cross-lingual dubbing that preserves the source speech's duration, speaker identity, and speaking speed. Existing methods mostly validate the translation performance on single direction, leaving the applicability of generalizing in multiple languages unexplored. The proposed method leverages both source and target language information that provides clear guidance to model language-specific semantics and pronunciations. Experimental results demonstrate that the proposed method shows comparable performance with existing approaches while effectively enabling training the translation model with multi-lingual data.

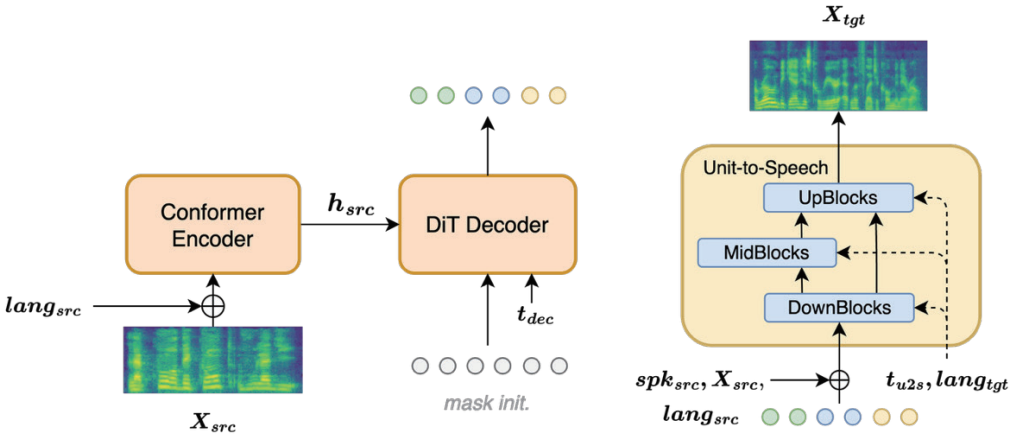

Figure 1. Overall architecture. (a) The source speech is encoded with language information to disambiguate the training process. (b) Synthesis of speech is conditioned with both source and target language information in different stages.

Ⅰ. Introduction

Recent advancements in textless speech-to-speech translation (S2ST) have paved the way for high-quality automatic dubbing systems [1]. The Dub-S2ST framework represents a significant step forward by generating translations that preserve key characteristics of the original speech, such as duration, speaker identity, and speaking speed [2]. This ability to generate time-aligned translations without manual post-processing, which can degrade quality, makes it highly suitable for dubbing applications.

Despite its promising results, a primary limitation of the existing Dub-S2ST model is its monolingual design. The model was trained and evaluated exclusively on a single language pair. This presents a scalability challenge for real-world dubbing scenarios, where multimedia content often needs to be translated from a diverse set of source languages into a common target language. Building and maintaining separate models for each language pair is inefficient and fails to leverage potential cross-lingual synergies.

To address this, we propose a multilingual extension of the framework designed for many-to-one speech dubbing. Our primary contribution is a method to incorporate language identity into the translation process using learnable language embeddings This approach enables a single model to be trained on a combined corpus of multiple language pairs, such as French-to-English and Spanish-to-English. The proposed method demonstrate that this multilingual extension can achieve translation quality comparable to dedicated single-language models while critically retaining the precise duration and speed control that are essential for seamless dubbing.

Ⅱ. Method

We extend the existing framework to a multi-lingual setting by introducing language conditional mechanisms into the speech encoding and Unit-to-Speech synthesis processes. To enable a single model to handle multiple source languages, we introduce learnable language embeddings. Each language present in the training corpus (e.g., French, Spanish, English) is assigned a unique, randomly initialized vector that is optimized during training.

The source speech mel-spectrogram (X_{src}) is first processed by the Conformer encoder to produce hidden representations (h_{src}). We position the fusion operation of source language information prior to the main encoding procedure by adding it to the mel-spectrogram. This explicitly informs the encoder about the linguistic context of the input, encouraging it to disentangle semantic content from language-specific acoustic properties and produce a more language-agnostic representation for translation.

During the synthesis of target speech based on the generated speech units, we first apply similar addition operation of the source language embedding. This reinforces the information potentially lost during the complex translation process. Subsequently, we condition the synthesis with the target language information via Adaptive Layer Normalization (AdaLN) [3]

$$h_{tgt} = \gamma(lang_{tgt}) * LayerNorm(h_{tgt}) + \beta(lang_{tgt}),$$

where h_{tgt}, $lang_{tgt}$, γ, β denotes target latent, target language embedding, and adaptive mean and variance, respectively. This operation effectively directs the latent distribution to favor the generation of target language by disambiguating against the source language.

We validate the effectiveness of the proposed method with CVSS-C dataset and utilize French and Spanish to English subset for multi-lingual translation. We compare the proposed method with the baseline Dub-S2ST model with French-to-English translation and present Spanish-to-English performance to evaluate the retention of performance when trained with multi-lingual translation dataset. We utilize ASR-BLEU for translation quality and DNSMOS for speech naturalness evaluation.

Translation pair	Method	ASR-BLEU	BLASER 2.0
fr-en	Dub-S2ST	24.16	3.839
	Ours	23.98	3.825
es-en	Dub-S2ST	23.51	3.811
	Ours	22.57	3.802

Table 1. Evaluation results on CVSS-C fr-en and es-en subsets. Dub-S2ST model is trained separately for the two datasets.

Ⅲ. Results

We demonstrate that the model trained with both French and Spanish to English datasets achieves strong performance across both language pairs. Based on ASR-BLEU and BLASER 2.0 being competitive with those of the individually trained baseline models, the proposed method shows effective knowledge transfer and generalization.

Ⅳ. Conclusion

In this paper, we propose a multilingual translation model for many-to-one speech dubbing. By conditioning the encoder and synthesis on source and target language embeddings, respectively, our model can process multiple source languages within a single, unified architecture. This work represents a significant step towards building a practical, high-quality automatic dubbing system capable of serving global multimedia platforms.

ACKNOWLEDGMENT

This work was supported by IITP-ITRC funded by the Korean Government (MSIT) under Grant IITP-2025-RS-2023-0025999

참 고 문 헌

[1] Federico M, Enyedi R, Barra-Chicote R, Giri R, Isik U, Krishnaswamy A, Sawaf H. “From speech-to-speech translation to automatic dubbing,” IWSLT 2020.

[2] Choi J, Kim J, Chung JS, “Dub-S2ST: Textless Speech-to-Speech Translation for Seamless Dubbing,” In Proc. EMNLP, 2025.

[3] Perez E, Strub F, De Vries H, Dumoulin V, Courville A. “Film: Visual reasoning with a general conditioning layer,” In Proc of AAAI, 2018.