Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Human-AI Interaction (Sakriani Sakti (Professor))		
Student ID	2411807	Submission date	2025 / 7 / 22
Name	JAN MEYER SARAGIH		
Thesis title	Enhancing Automatic Dubbing through Translation Diversification and Synchrony-Conscious Re-Ranking		

Abstract

Translation plays a vital role in enabling multilingual communication across domains such as international media, education, and live events. Recent advances in translation technology have spanned a range of modalities, including text-to-text, speech-to-text, and speech-to-speech translation. Among these, simultaneous translation—designed to approximate real-time human interpretation—has received considerable attention due to its utility in latency-sensitive applications.

In contrast, dubbing—another important form of speech-to-speech translation—remains relatively underexplored. Dubbing replaces the original spoken content with translated speech in a different language, typically for audiovisual media such as films, documentaries, or instructional videos . Unlike simultaneous translation, dubbing prioritizes temporal synchronization with the source audio. This involves not only matching total utterance duration, but also preserving pause timing, rhythm of phrasing, and segment alignment—all while maintaining semantic fidelity and fluency.

Despite its practical importance, dubbing has received limited attention in the machine translation community. Some recent approaches attempt to impose duration constraints during generation—using auxiliary duration models, token-level length tags, or hard counters—but these often degrade in translation quality. Others, introduce large language models (LLMs) for paraphrasing guided by prosodic segmentation, but these methods rely on a single input translation and do not explore a broader space of alternatives.

At the same time, professional dubbing workflows frequently prioritize natural-sounding translations, even when perfect duration alignment is not achieved. This highlights the need for systems that can flexibly balance semantic accuracy and temporal synchrony by exploring multiple translation variants.

In this thesis, a multi-candidate framework for dubbing-aware translation is introduced to address the previously mentioned limitations. The contribution in this thesis includes the introduction of three main components. First, the generate of N-best list of translation candidates from machine translation system, expanding the semantic and prosodic search space without modifying the base model. Second, the application of in-context paraphrasing with large language models to generate diverse rephrasings for each MT candidate. These paraphrases aim to preserve meaning while introducing lexical and structural variation, enabling a wider space of alignment possibilities. Third, the implementation of synchrony-aware re-ranking using automatic speech-only metrics to select the most temporally aligned output. This step enforces alignment without requiring video or retraining of MT/TTS models.