Beyond One-Shot Dubbing: Leveraging N-Best Translation and Prompted Paraphrasing with Synchrony-Aware Re-Ranking

Jan Meyer Saragih, Faisal Mehmood, and Sakriani Sakti
Nara Institute of Science and Technology, Japan
E-mail: jan.meyersaragih.jn9@naist.ac.jp, mehmood.faisal@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—Dubbing—translating and synchronizing spoken content across languages—requires not only semantic accuracy but also precise alignment with the timing and rhythm of the original speech. While recent work has explored paraphrasing with large language models (LLMs) for dubbing, most approaches rely on a single translation and a one-shot paraphrasing pass, limiting the space of alternatives and often resulting in suboptimal alignment. We propose a multi-candidate framework that leverages N-best machine translation outputs and in-context paraphrasing with LLMs to generate diverse candidates that preserve meaning while varying in structure and length. This expanded candidate space enables better exploration of alignment options without modifying the underlying MT or TTS systems. To select the best output, we introduce synchrony-aware reranking based on speech-only alignment metrics—specifically Speech Length Compliance and Speech Overlap—which quantify temporal alignment between the generated and reference speech without requiring video input or retraining. Experiments on German-to-English dubbing using the FLEURS dataset and the IWSLT 2024 Dubbing Challenge test set show that our approach improves temporal synchrony and perceived naturalness. These results highlight the value of combining translation diversity, in-context paraphrasing, and synchrony-based selection for high-quality automatic dubbing.**

I. Introduction

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 [1], speech-to-text [2], and speech-to-speech translation [3]. Among these, simultaneous translation—designed to approximate real-time human interpretation—has received considerable attention due to its utility in latency-sensitive applications [4].

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 [5]. 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 [6].

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 [7]. Others, such as [8], introduce large language models (LLMs) for one-shot 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 [9]. This highlights the need for systems that can flexibly balance semantic accuracy and temporal synchrony by exploring multiple translation variants.

In this paper, we propose a multi-candidate framework for dubbing-aware translation that addresses these limitations. Our approach introduces three main components:

- **First**, we generate an *N-best list of translation candidates* from a machine translation system, expanding the semantic and prosodic search space without modifying the base model.
- **Second**, we apply *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**, we perform *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.

Together, these components form a modular, language-agnostic pipeline for high-quality automatic dubbing. Our framework operates entirely with off-the-shelf translation and synthesis systems, and enables alignment-sensitive selection through automatic evaluation and re-ranking.

II. Related Work

N-best translation and re-ranking have long been core components of statistical machine translation (SMT), where multiple candidate hypotheses are generated and ranked using

979-8-3315-7206-8/25/$31.00

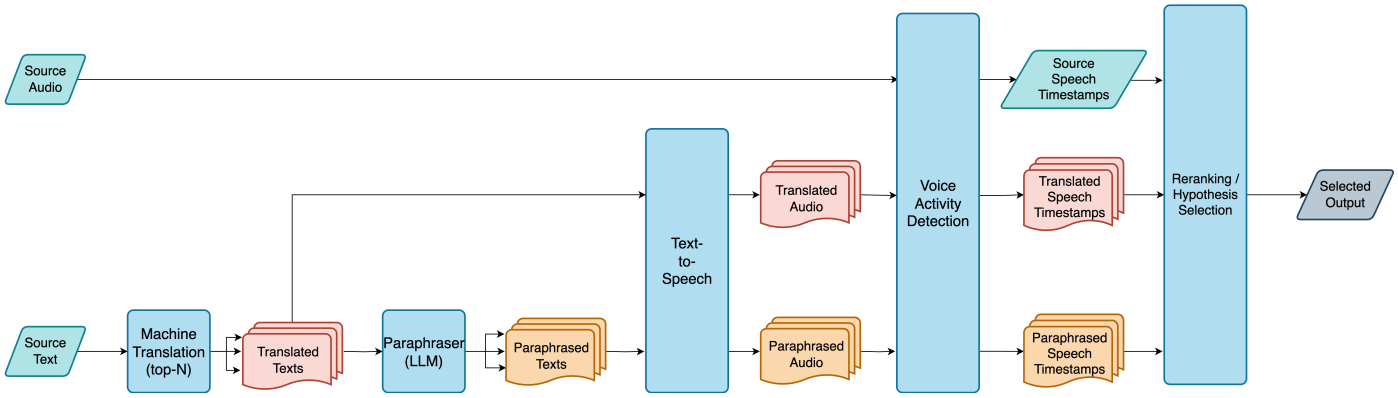

Fig. 1. System Overview

alignment features, language models, or classifier-based scoring functions [10]. In the neural MT era, N-best lists remain valuable for tasks such as domain adaptation and post-editing [11], but the focus remains largely on improving semantic fidelity and fluency, without attention to temporal or prosodic properties.

Dubbing-aware translation, in contrast, introduces unique temporal constraints. The aim is not only to produce accurate translations, but also to match the timing and rhythm of the original speech. Recent efforts to address this include generation-time constraints—such as duration prediction models, length tags, and token-level pacing controls [7], [12] and post-generation methods like paraphrasing a single translation via large language models (LLMs) conditioned on prosodic segmentation [8]. While promising, these methods operate in a one-shot fashion and fail to fully explore the space of paraphrastic alternatives that could yield better synchrony.

Our work extends the traditional N-best re-ranking paradigm to the dubbing setting, incorporating **in-context paraphrasing** with LLMs to generate diverse rephrasings of translation candidates. Rather than explicitly encoding timing constraints in the generation step, we rely on **synchrony-aware re-ranking**—using speech-only alignment metrics such as *Speech Length Compliance* and *Speech Overlap*—to identify the most temporally compatible output. This allows us to optimize alignment in a post-hoc, model-agnostic fashion. These speech-based metrics offer a practical alternative to commonly used alignment evaluation methods like *Lip Sync Error Distance (LSE-D)* or *character error rate (CER)*, which often require video input or duration-predictive models. In contrast, our pipeline evaluates synchrony purely from speech audio and remains fully compatible with off-the-shelf MT and TTS systems. To our knowledge, this is the first work to integrate N-best MT generation, LLM-based paraphrasing, and automatic alignment-driven re-ranking into a unified framework for dubbing-aware translation.

III. System Description

We present a multi-candidate framework for dubbing-aware translation that integrates translation diversity, in-context paraphrasing, speech synthesis, and synchrony-aware re-ranking. As illustrated in Figure 1, the system comprises five core components, each contributing to the generation of semantically accurate and temporally aligned dubbed speech.

A. Machine Translation: Generating N-best Translations

The pipeline begins with a machine translation component that generates an N-best list of translation hypotheses, introducing semantic diversity into the system. This diversity is critical, as different formulations of the same utterance may naturally vary in their timing characteristics and alignment potential. We use SeamlessM4T [13], a high-quality multilingual model, to produce multiple plausible translations for each input utterance. These serve as the base candidates for downstream alignment-aware paraphrasing.

B. Paraphrasing: In-Context Prompting with Semantic and Synchrony Objectives

Each MT candidate is passed through a paraphrasing module using in-context prompting with a large language model (LLM). We employ Qwen3 [14], prompting it to generate paraphrases that preserve meaning while enhancing either semantic diversity or temporal compatibility with the source speech. Our prompting strategies fall into three categories:

- **Semantic-Oriented:** Prompts that encourage natural, detailed rephrasings while maintaining the original meaning.
- **Semantic + Structural:** Prompts that explicitly encourage lexical and structural variation, offering diverse rewrites without enforcing alignment constraints.
- **Synchrony-Oriented:** Prompts that guide the LLM to match the character length of the source utterance, serving as a proxy for duration alignment in TTS.

By leveraging both semantic and synchrony-aware prompts, our framework generates a diverse candidate pool with controllable properties. This allows downstream re-ranking to select outputs that best balance fluency and temporal alignment—without modifying the MT or TTS components.

C. Text-to-Speech: Synthesizing Candidate Speech

Each paraphrased text is then passed through a text-to-speech system to generate audio representations of the candidate utterances. We use MeloTTS [15], chosen for its ability

to produce expressive, natural-sounding speech with accurate timing characteristics. The synthesized audio allows us to evaluate how each paraphrase translates into real-world timing and whether it aligns well with the source speech. This component provides the acoustic basis for temporal evaluation.

D. Voice Activity Detection: Extracting Temporal Structure

To analyze timing, we extract voice activity segments from both the original reference audio and each synthesized candidate using Silero VAD [16]. This tool detects speech-active intervals, enabling us to measure the start and end times of speech segments in a lightweight and language-independent way. By comparing the segment durations and pause placements between the source and each candidate, we can quantify their temporal alignment without requiring access to video or lip-sync data.

E. Re-ranking: Synchrony-Aware Candidate Selection

The final stage of our pipeline is a synchrony-aware re-ranking module that selects the most temporally aligned candidate based on automatic speech-level analysis. The process involves:

- Combining the N-best MT outputs with their paraphrased variants generated via in-context prompting.
- Filtering out semantically weak candidates using lexical or embedding-based similarity measures.
- Synthesizing speech for each remaining candidate using a TTS system, and extracting timestamps via voice activity detection (VAD).
- Computing alignment scores using *Speech Length Compliance (SLC)* and *Speech Overlap*.
- Selecting the candidate with the highest alignment score for the final output.

This speech-only synchrony-aware re-ranking strategy enables efficient selection without requiring video input or retraining of MT/TTS systems.

IV. Experimental Setup

A. Datasets

We conduct our experiments on two evaluation datasets: **FLEURS** and the **IWSLT 2024 Dubbing Test Set**, both of which provide German source speech aligned with English target speech and text—enabling evaluation of both translation quality and temporal alignment.

- **FLEURS** [17]
  A multilingual speech dataset with high-quality, sentence-aligned audio in over 100 languages. We used the German and English set. This dataset has both German and English audio, but we opted to use the German source speech to maintain consistency with the other dataset used in our experiments.
- **IWSLT 2024 Dubbing Test Set** [18]
  A professionally re-recorded subset of the CoVoST-2 dataset released for the IWSLT 2024 Dubbing Challenge. This dataset was recorded using German speakers and includes aligned English references speech/text. It also contains video, which we do not use in our experiments.

B. Baseline

We compare our system against the official baseline from the IWSLT 2024 Dubbing Challenge [19]. The baseline includes a machine translation model augmented with phoneme-level duration prediction, allowing it to optimize alignment between translation output and the timing of source speech during decoding. This tightly integrated approach serves as a strong reference point for evaluating how well our modular, re-ranking-based method can achieve comparable or better temporal alignment and translation quality without retraining the MT or TTS components.

C. Evaluation Metrics

We evaluate system performance across two main dimensions: **translation quality** and **temporal alignment**.

Translation Quality: Assessed using standard lexical and semantic metrics:

- **BLEU** [20] – measures word-level overlap with the reference.
- **XCOMET** [21] and **CometKiwi** [22] – neural metrics for evaluating semantic adequacy and fluency.

Temporal Alignment: Evaluated using speech-only metrics:

- **Speech Length Compliance (SLC)** [23] – compares total duration of source and generated speech.
- **Speech Overlap (SO)** [24] – measures overlap across predefined speech segments.

D. Focus of Analysis

Our experiments are designed to answer two key questions about the structure and effectiveness of our proposed framework:

- **Impact of Candidate Diversity:** We begin by comparing our system against the IWSLT 2024 baseline, which applies duration control within the MT decoder. In contrast, our approach generates an N-best list of diverse translation outputs (without paraphrasing) and applies synchrony-based re-ranking to select the most temporally aligned candidate. We also compare multiple re-ranking strategies— SLC versus SO—to assess which metric better captures alignment quality without modifying MT or TTS components.
- **Investigating In-Context Paraphrasing:** We then evaluate whether synchrony-aware paraphrasing via large language models (LLMs) further improves alignment. Specifically, we test prompt strategies that emphasize semantic enrichment (e.g., longer paraphrases), structural variation (e.g., rewording and reordering), and temporal control (e.g., character-length matching). In order to keep the comparison fair, the amount of candidates is made to be the same as N-best list candidates.

V. Experiment Results and Discussion

A. Impact of Candidate Diversity

We evaluate the effect of candidate diversity by comparing three configurations: the official **Baseline** system from the IWSLT 2024 Dubbing Challenge, the **Top-1 MT** output from

TABLE I
IMPACT OF CANDIDATE DIVERSITY ON TRANSLATION QUALITY AND TEMPORAL ALIGNMENT ACROSS SYSTEMS AND DATASETS

System	Re-ranking	Dataset	BLEU	COMET-MT	COMET-QE	SLC	SO
Baseline	–	FLEURS	14.16	0.531	0.361	0.750	0.430
		IWSLT Subset 1	28.85	0.830	0.534	0.841	0.686
		IWSLT Subset 2	19.20	0.605	0.422	0.847	0.595
Top-1 MT	–	FLEURS	42.31	0.965	0.756	0.778	0.371
		IWSLT Subset 1	50.92	0.985	0.773	0.678	0.528
		IWSLT Subset 2	50.78	0.973	0.764	0.706	0.058
Top-5 MT	Sorted by SLC	FLEURS	37.32	0.963	0.755	0.805	0.388
		IWSLT Subset 1	39.56	0.982	0.761	0.736	0.576
		IWSLT Subset 2	47.48	0.969	0.762	0.749	0.040
	Sorted by SO	FLEURS	38.56	0.967	0.757	0.792	0.415
		IWSLT Subset 1	41.33	0.984	0.763	0.728	0.577
		IWSLT Subset 2	49.57	0.977	0.767	0.712	0.087

SLC = Speech Length Compliance, **SO** = Speech Overlap (Segment-based). Systems are evaluated on three datasets using standard translation metrics (BLEU, COMET) and temporal alignment metrics.

our neural machine translation (NMT) model, and the proposed **Top-5 MT + Re-Ranking** framework. Table I presents results across three datasets: FLEURS, IWSLT Subset 1, and IWSLT Subset 2.

The **Baseline** system incorporates duration modeling directly into the MT decoder, resulting in relatively strong alignment performance—e.g., Speech Length Compliance (SLC) scores of 0.847 and 0.841 on the IWSLT subsets, and moderate Speech Overlap (SO) scores of 0.686 and 0.595. However, translation quality is substantially lower, with BLEU scores falling to 14.16 on FLEURS and COMET-MT scores below 0.83, indicating limited semantic fidelity.

The **Top-1 MT** configuration, which selects the best output from a standard neural MT model without duration control, significantly improves translation quality. BLEU exceeds 50 and COMET-MT surpasses 0.97 on the IWSLT subsets. However, this gain comes at the cost of alignment: SO drops sharply—e.g., to 0.058 on IWSLT Subset 2—and SLC scores also decline, highlighting the trade-off between fluency and temporal synchrony.

To mitigate this, our **Top-5 MT + Re-Ranking** approach broadens the candidate space by generating N-best translations. These are paraphrased using in-context prompts and then re-ranked using alignment metrics. While BLEU scores are slightly lower than Top-1—due to greater lexical variation and less exact overlap with the reference—the COMET-MT scores remain high ($>$0.96), showing that semantic fidelity is preserved. This is consistent with our objective of producing temporally aligned, fluent translations rather than maximizing reference overlap.

We compare two re-ranking strategies: **SLC** and **SO** (segment-based). Both outperform Top-1 in terms of alignment, but SO-based re-ranking generally achieves slightly better results. For example, on FLEURS, SO increases from 0.371 (Top-1) to 0.415 (SO-based), while SLC rises from 0.778 to 0.805 (SLC-based). These trends suggest that segment-level SO, which captures fine-grained temporal overlap, is a more precise indicator of speech alignment than duration alone.

In summary, these results show that:

- Translation quality and temporal alignment are often at odds, motivating the need for multi-candidate exploration.
- Re-ranking diverse candidates using speech-based alignment metrics can restore synchrony without sacrificing semantic quality.
- Segment-based SO is more sensitive to alignment quality than global duration measures like SLC, though both are complementary.

B. Investigating In-Context Paraphrasing

To assess the role of in-context paraphrasing in dubbing-aware translation, we compare three prompting strategies applied to N-best MT outputs, as summarized in Table II. Each prompt targets a different balance of semantic preservation, structural variation, and timing control. Their impact on translation quality and alignment is reported in Table III, with candidates re-ranked by either **Speech Length Compliance (SLC)** or **Speech Overlap (SO)**.

Prompt 1 (Semantic-Oriented) consistently produces high COMET-MT scores (e.g., 0.978 on IWSLT Subset 1) and natural fluency. However, its improvements in alignment are limited. While SLC remains strong (e.g., 0.925 on Subset 2), SO values are modest across datasets, indicating that paraphrasing focused solely on meaning enrichment is insufficient for precise temporal synchronization.

Prompt 3 (Synchrony-Oriented) aims to enforce timing control via character-length constraints. However, it underperforms in both semantic and alignment metrics. BLEU and COMET-MT scores are generally lower than the other prompts, and SO remains similar or worse (e.g., SO = 0.047 on Subset 2). These results suggest that rigid surface-level control can reduce fluency without consistently enhancing alignment.

Prompt 2 (Semantic + Structural) yields the most favorable balance. It improves segment-level SO while maintaining high COMET-MT (e.g., 0.976 on Subset 2). SO values surpass those of the other prompts on most datasets—e.g., 0.769 on Subset 1 and 0.197 on Subset 2—demonstrating that structural variation during paraphrasing provides flexibility that benefits alignment without compromising meaning.

Re-ranking by SO consistently leads to stronger alignment compared to SLC-based ranking, across all prompt types. For

TABLE II
IN-CONTEXT PARAPHRASING

Prompt	Category	Prompt Description
Prompt 1	Semantic-Oriented	You are a professional paraphraser. You are given a German sentence and its English translation. Paraphrase the given English translation. Keep the meaning the same but rephrase it naturally with more detail. Output only the new sentence.
Prompt 2	Semantic + Structural	You are an expert paraphrasing assistant. Now you are given the original German text and its English translation. Your goal is to rewrite the English sentence in a way that keeps the original meaning but changes the structure and vocabulary as much as possible. Use natural, fluent English. Avoid repeating key phrases from the original.
Prompt 3	Synchrony-Oriented	You are an expert paraphrasing assistant. You are given a German sentence and its English translation. Your task is to revise the provided English translation so that its number of characters closely matches that of the original German sentence. Maintain the original meaning and ensure the English remains natural and fluent. Output only the revised English sentence.

TABLE III
IMPACT OF IN-CONTEXT PARAPHRASING ON TRANSLATION QUALITY AND TEMPORAL ALIGNMENT ACROSS SYSTEMS AND DATASETS

System	Re-ranking	Dataset	BLEU	COMET-MT	COMET-QE	SLC	SO Segment
Paraphrase Prompt 1	Sorted by SLC	FLEURS	27.76	0.961	0.745	0.909	0.437
		IWSLT Subset 1	23.00	0.971	0.717	0.856	0.664
		IWSLT Subset 2	18.30	0.952	0.724	0.925	0.046
	Sorted by SO	FLEURS	31.73	0.965	0.748	0.826	0.625
		IWSLT Subset 1	25.77	0.973	0.724	0.819	0.772
		IWSLT Subset 2	39.58	0.971	0.756	0.743	0.194
Paraphrase Prompt 2	Sorted by SLC	FLEURS	40.95	0.970	0.760	0.896	0.477
		IWSLT Subset 1	40.04	0.981	0.742	0.851	0.675
		IWSLT Subset 2	44.90	0.969	0.759	0.865	0.054
	Sorted by SO	FLEURS	42.28	0.971	0.761	0.812	0.625
		IWSLT Subset 1	41.74	0.982	0.749	0.810	0.769
		IWSLT Subset 2	49.11	0.976	0.763	0.705	0.197
Paraphrase Prompt 3	Sorted by SLC	FLEURS	41.23	0.969	0.761	0.901	0.454
		IWSLT Subset 1	20.76	0.977	0.726	0.852	0.685
		IWSLT Subset 2	46.00	0.968	0.756	0.858	0.047
	Sorted by SO	FLEURS	41.91	0.971	0.761	0.805	0.620
		IWSLT Subset 1	24.12	0.979	0.738	0.817	0.776
		IWSLT Subset 2	47.57	0.974	0.762	0.715	0.202

SLC = Speech Length Compliance, **SO** = Speech Overlap. Systems are evaluated on three datasets using standard translation metrics (BLEU, COMET) and temporal alignment metrics.

example, for Prompt 2 on Subset 1, SO rises from 0.675 (SLC) to 0.769 (SO), and on Subset 2, from 0.054 to 0.197. These results confirm that segment-based SO offers a more fine-grained and effective alignment signal than global duration matching.

Compared to Top-5 MT without paraphrasing (Table I), paraphrasing significantly enhances alignment while maintaining semantic quality. For instance, Prompt 2 with SO-based re-ranking improves SO on Subset 2 from 0.040 (Top-5 MT) to 0.197, while keeping COMET-MT above 0.97. BLEU scores decrease slightly due to lexical variability, but this trade-off is acceptable in dubbing applications prioritizing fluency and synchrony.

In summary, we find that:

- **Prompt 2** (semantic + structural) offers the best trade-off between alignment and translation quality, outperforming **Prompt 1** (strong fluency, weaker alignment) and **Prompt 3** (minimal alignment gain, often at the cost of fluency).
- **SO-based re-ranking** is more effective than SLC for selecting well-aligned candidates.
- **Paraphrasing significantly outperforms** the baseline Top-5 MT setting in synchrony, while preserving translation quality.

VI. CONCLUSIONS

We proposed a flexible, modular framework for dubbing-aware translation that goes beyond one-shot generation by integrating N-best MT decoding, in-context paraphrasing, and speech-based re-ranking. Our experiments demonstrate that candidate diversity plays a central role in improving alignment: selecting among Top-5 MT outputs already improves synchrony compared to single-best outputs, even without modifying the MT or TTS components. Building on this, we show that in-context paraphrasing—particularly with semantic + structural prompts—enables the model to discover temporally compatible alternatives without sacrificing meaning. While rigid synchrony-oriented prompts (e.g., character constraints) offered limited benefits, structural variation improved alignment with natural fluency. Re-ranking with segment-based Speech Overlap (SO) consistently outperformed global duration metrics like SLC. These findings support a candidate-centric view of dubbing: rather than tightly constraining generation, we encourage diversity and optimize selection based on alignment metrics. Overall, our results highlight the effectiveness of combining translation diversity, controlled paraphrasing, and synchrony-aware selection—paving the way for more adaptive and scalable dubbing pipelines.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681, as well as JST NEXUS (JPMJNX25C1).

REFERENCES

[1] M. Cui, P. Gao, W. Liu, J. Luan, and B. Wang, "Multilingual machine translation with open large language models at practical scale: An empirical study," in *Proc. NAACL-HLT (Volume 1: Long Papers)*, Association for Computational Linguistics, 2025, pp. 5420–5443.

[2] T. K. Lam, M. Gaido, S. Papi, L. Bentivogli, and B. Haddow, "Prepending or cross-attention for speech-to-text? an empirical comparison," in *Proc. NAACL-HLT (Volume 1: Long Papers)*, Association for Computational Linguistics, 2025, pp. 2994–3006.

[3] T. Labiausse, L. Mazaré, E. Grave, P. Pérez, A. Défossez, and N. Zeghidour, "High-fidelity simultaneous speech-to-speech translation," *arXiv preprint arXiv:2502.03382*, 2025.

[4] Y. Ko, R. Fukuda, Y. Nishikawa, *et al.*, "NAIST simultaneous speech translation system for IWSLT 2024," in *Proc. IWSLT 2024*, Association for Computational Linguistics, 2024, pp. 170–182.

[5] K. Sung-Bin, J. Choi, P. Peng, J. S. Chung, T.-H. Oh, and D. Harwath, "Voicecraft-dub: Automated video dubbing with neural codec language models," *arXiv preprint arXiv:2504.02386*, 2025.

[6] M. Federico, Y. Virkar, R. Enyedi, and R. Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," in *Interspeech 2020*, 2020, pp. 1481–1485.

[7] P. Pal, B. Thompson, Y. Virkar, P. Mathur, A. Chronopoulou, and M. Federico, "Improving isochronous machine translation with target factors and auxiliary counters," in *Proc. Interspeech 2023*, 2023, pp. 37–41.

[8] Y. Li, J. Guo, M. Zhang, *et al.*, "Pause-aware automatic dubbing using LLM and voice cloning," in *Proc. IWSLT 2024*, Association for Computational Linguistics, 2024, pp. 12–16.

[9] W. Brannon, Y. Virkar, and B. Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 419–435, 2023.

[10] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[11] M. Freitag, D. Grangier, and I. Caswell, "BLEU might be guilty but references are not innocent," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 61–71.

[12] Z. Rao, H. Shang, J. Yang, *et al.*, "Length-aware NMT and adaptive duration for automatic dubbing," in *Proc. IWSLT 2023*, Association for Computational Linguistics, 2023, pp. 138–143.

[13] L. Barrault, Y.-A. Chung, M. C. Meglioli, *et al.*, "Seamlessm4t: Massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[14] Y. Zhang, M. Li, D. Long, *et al.*, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," *arXiv preprint arXiv:2506.05176*, 2025.

[15] W. Zhao, X. Yu, and Z. Qin, *Melotts: High-quality multi-lingual multi-accent text-to-speech*, 2023.

[16] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, 2024.

[17] A. Conneau, M. Ma, S. Khanuja, *et al.*, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 798–805.

[18] E. Salesky, M. Federico, and M. Carpuat, Eds., *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, Association for Computational Linguistics, 2024.

[19] P. Pal, B. Thompson, Y. Virkar, P. Mathur, A. Chronopoulou, and M. Federico, "Improving isochronous machine translation with target factors and auxiliary counters," in *Proc. Interspeech 2023*, 2023, pp. 37–41.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[21] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, and A. F. T. Martins, "Xcomet: Transparent machine translation evaluation through fine-grained error detection," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 979–995, 2024.

[22] R. Rei, N. M. Guerreiro, J. Pombal, *et al.*, "Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task," in *Proceedings of the Eighth Conference on Machine Translation*, Association for Computational Linguistics, 2023, pp. 841–848.

[23] Y. Wu, J. Guo, X. Tan, *et al.*, "Videodubber: Machine translation with speech-aware length control for video dubbing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 13 772–13 779.

[24] P. Pal, B. Thompson, Y. Virkar, P. Mathur, A. Chronopoulou, and M. Federico, "Improving isochronous machine translation with target factors and auxiliary counters," in *Proc. Interspeech 2023*, 2023, pp. 37–41.