End-to-End Multilingual Automatic Dubbing via Duration-based Translation with Large Language Models

Hyun-Sik Won[*], **DongJin Jeong**[*], **HyunKyu Choi**[*], **JinWon Kim**[*†]
ESTsoft
{abugda, jdjin3000, choihk, jw93}@estsoft.com

Abstract

Automatic dubbing (AD) aims to replace the original speech in a video with translated speech that maintains precise temporal alignment (isochrony). Achieving natural synchronization between dubbed speech and visual content remains challenging due to variations in speech durations across languages. To address this, we propose an end-to-end AD framework that leverages large language models (LLMs) to integrate translation and timing control seamlessly. At the core of our framework lies Duration-based Translation (*DT*), a method that dynamically predicts the optimal phoneme count based on source speech duration and iteratively adjusts the translation length accordingly. Our experiments on English, Spanish, and Korean language pairs demonstrate that our approach substantially improves speech overlap—achieving up to 24% relative gains compared to translations without explicit length constraints—while maintaining competitive translation quality measured by COMET scores. Furthermore, our framework does not require language-specific tuning, ensuring practicality for multilingual dubbing scenarios. We also provide an online demo[1] and a demo video[2].

1 Introduction

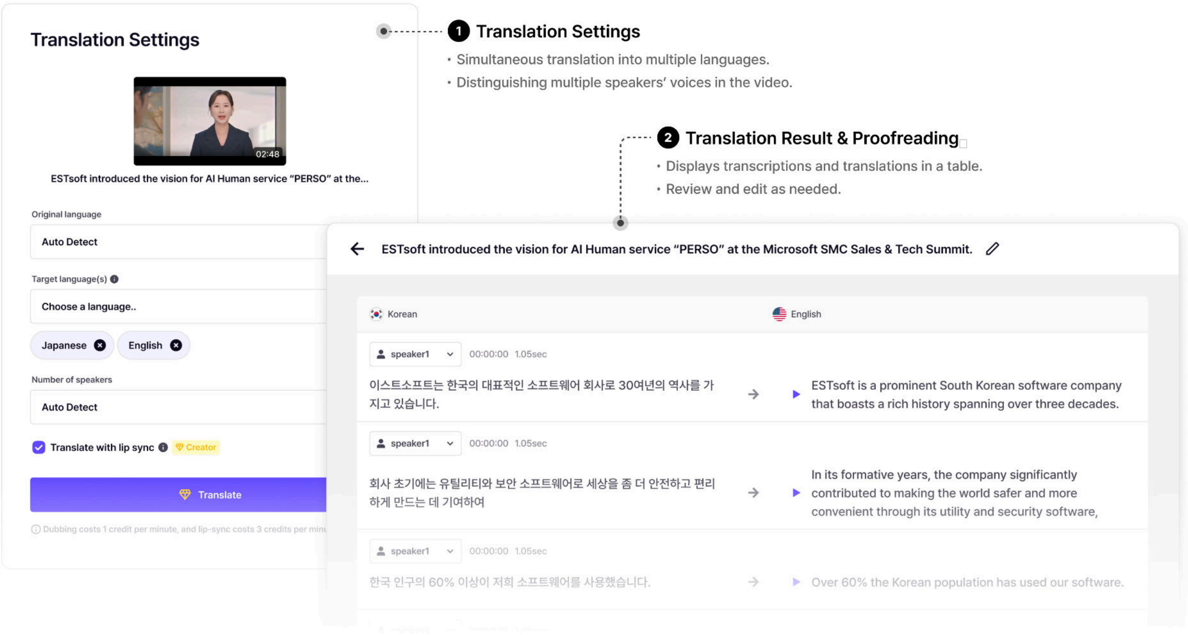
Automatic dubbing (AD) aims to translate the spoken content of a video (e.g., films or TV shows) into another language and replace the source speech with a translated voice track, while preserving natural timing and synchronization (Virkar et al., 2022). A challenge in AD is isochrony, which requires the translated speech to be temporally aligned with the source speaker's mouth movements and pauses (Chaume, 2008). To achieve isochrony, the speech–pause pattern of the source must be preserved in the translation, with corresponding segments in the target speech maintaining the same temporal alignment as the original. If the translated speech does not achieve isochrony, the result will look and sound unnatural because the audio will be out of sync with on-screen speech or lip movements. Most AD systems (Rao et al., 2023; Wu et al., 2023) follow a cascade pipeline of Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) for synthesis. However, the NMT module is focused on optimizing for lexical accuracy, not for temporal alignment, resulting in translations that are mismatched with the source speech durations.

To align translations with the temporal structure of source speech, researchers have proposed various approaches. The study (Öktem et al., 2019) has proposed adjusting the speaking rate of TTS-generated speech to match the duration of the source speech. However, excessively manipulating the speaking rate can distort the resulting speech, making it sound unnatural (Wu et al., 2023). Other studies (Lakew et al., 2022; Tam et al., 2022; Rao et al., 2023) have attempted to match translation lengths to source text lengths, aiming to achieve similar spoken durations. Nevertheless, equal character or word counts do not necessarily lead to similar speech durations, as speech rates vary depending on both the language and the particular speaker. Further studies (Chronopoulou et al., 2023; Pal et al., 2023; Wu et al., 2023) incorporate phoneme-duration prediction and length-control techniques, which improve isochrony but often add complexity and are tailored to specific language pairs, limiting scalability for real-world multilingual media.

In this work, we propose an end-to-end automatic dubbing framework that incorporates Duration-based Translation (*DT*), a translation module designed to dynamically adjust the length of translated text based on source speech durations,

[*] Equal contribution.
[†] Project lead; Current affiliation Hyundai AutoEver.
[1] https://perso.ai/
[2] https://youtu.be/N24pI4bsIfc

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 515–521
November 4-9, 2025 ©2025 Association for Computational Linguistics



Translation Settings

ESTsoft introduced the vision for AI Human service “PERSO” at the...

Original language: Auto Detect

Target language(s): Choose a language..

Japanese ☒ English ☒

Number of speakers: Auto Detect

☑ Translate with lip sync

Translate

Dubbing costs 1 credit per minute, and lip-sync costs 3 credits per minu

1 Translation Settings
- Simultaneous translation into multiple languages.
- Distinguishing multiple speakers’ voices in the video.

2 Translation Result & Proofreading
- Displays transcriptions and translations in a table.
- Review and edit as needed.

ESTsoft introduced the vision for AI Human service “PERSO” at the Microsoft SMC Sales & Tech Summit.

Korean	English
speaker1 00:00:00 1.05sec 이스트소프트는 한국의 대표적인 소프트웨어 회사로 30여년의 역사를 가지고 있습니다.	ESTsoft is a prominent South Korean software company that boasts a rich history spanning over three decades.
speaker1 00:00:00 1.05sec 회사 초기에는 유틸리티와 보안 소프트웨어로 세상을 좀 더 안전하고 편리하게 만드는 데 기여하여	In its formative years, the company significantly contributed to making the world safer and more convenient through its utility and security software,
speaker1 00:00:00 1.05sec 한국 인구의 60% 이상이 저희 소프트웨어를 사용했습니다.	Over 60% the Korean population has used our software.

Figure 1: The interface comprises two main sections: (A) The translation configuration panel provides options for selecting target languages. Users can specify the number of speakers and enable lip-synchronization features through this panel. (B) The subtitle presentation panel displays the original text along with its translations. This panel allows for review and editing of translated content.

achieving accurate synchronization. To satisfy this constraint, our framework iteratively refines the translation by dynamically shortening or lengthening phrases until it meets the desired phoneme count.

This approach ensures translations to be spoken at a natural pace, minimizing the need for any post-processing or TTS speed manipulation. Importantly, our framework supports multilingual scenarios without language-specific tuning, and we showcase its capabilities with a real-time demonstration. The main contributions of this study are summarized as follows:

- We propose a novel end-to-end AD framework that leverages the broad linguistic capability of large language models (LLMs) to integrate translation and time constraint, ensuring natural synchronization without the need for extensive post-processing or TTS speed manipulation. (Figure 1)

- We introduce a phoneme count predictor that estimates the optimal number of phonemes for each translated segment, enabling dynamic length adjustments to maintain seamless timing with the source speech.

- We validate our approach through extensive experiments and demonstrate that our approach significantly improves temporal alignment while maintaining competitive translation quality.

2 Related Work

Isometric Translation via Length Constraint Several studies have proposed “isometric translation”, an approach that controls the translation length to closely match the source text. Lakew et al. (2022) address the challenge of controlling translation length for automatic dubbing by introducing a self-learning approach that trains MT models to generate translations within ±10% of the source text length. Their method enables direct generation of length-appropriate translations, eliminating the traditional two-step process of generating multiple hypotheses and then reranking them based on length criteria.

Tam et al. (2022) effectively advance isometric translation research through their methods. They explore two approaches: an implicit strategy that inserts pause markers directly into the text, and an explicit strategy that employs length-dependent positional embeddings based on character count ratios. Their explicit approach controls output length at the phrase level by managing character counts between pauses, effectively achieving isometric translation

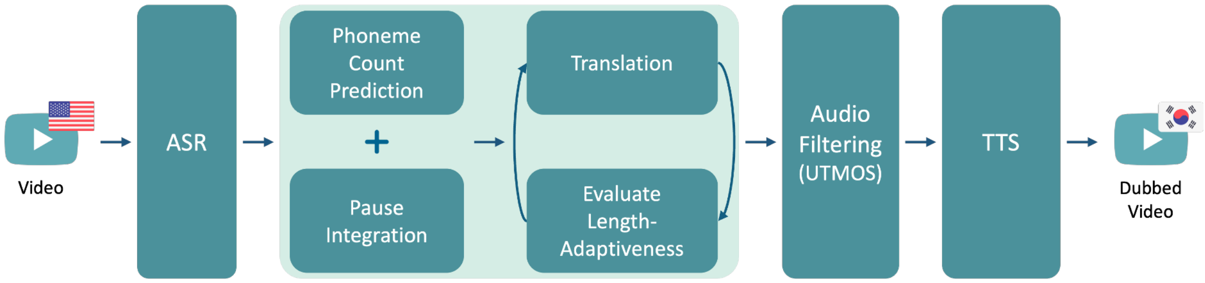


Figure 2: Overview of the proposed AD framework. In the framework, ASR transcription initiates the pipeline, enhanced via (1) phoneme count prediction, which utilizes original speech duration to estimate appropriate phoneme length and (2) pause integration, which captures temporal dynamics of the source speech. The enhanced inputs feed into a translation process, where translation and length-adaptiveness evaluation form an iterative feedback loop for proper synchronization. UTMOS filtering selects quality samples for voice cloning, then TTS synthesis creates natural-sounding dubbed output.

in segments rather than just at the sentence level. Rao et al. (2023) propose multiple target-to-source length-ratio labels (e.g., XSHORTER, SHORTER, EQUAL, LONGER, XLONGER). By training the model to generate translations aligned with these discrete labels, they achieve flexible length control that addresses various isochrony demands in real-world AD scenarios.

Translation with Duration Modeling An alternative line of work integrates speech timing constraints directly into the translation model. Chronopoulou et al. (2023) propose a framework that interleaves phonemes with their corresponding durations to jointly learn translation and duration prediction. Extending this framework, Pal et al. (2023) propose simultaneously predicting the sequence and duration of each phoneme, employing auxiliary indicators to track the remaining timing. Similarly, Wu et al. (2023) proposes VIDEODUBBER, which incorporates a duration predictor, duration-aware positional embeddings, and a special pause token in the decoder that enables fine-grained control over speech rhythm and synchronization. By directly predicting the speaking duration of each target token, these approaches enable a closer synchronization with the source speech compared to methods that rely on isometric way.

Toward a General-Purpose Solution Although previous studies demonstrate that employing text-level length constraints or integrating duration modeling can effectively address isochrony, the language-specific engineering burden remains substantial. To address this limitation, Li et al. (2024) propose an LLM-based approach that generates multiple translation candidates, synthesizes speech for each candidate, and selects the optimal dub based on audio-based evaluation metrics. However, this post-hoc approach cannot predict the synthesized speech duration before generating the audio. To overcome this limitation, we propose a phoneme count predictor that eliminates the need for iterative TTS synthesis.

3 Methodology

In this section, we detail the architecture of our end-to-end automatic dubbing framework, focusing on the design of its key component, Duration-based Translation (*DT*), which ensures accurate temporal synchronization. Figure 2 illustrates the overall pipeline, consisting of three sequential modules: Speech-to-Text (STT), Neural Machine Translation (NMT), and Text-to-Speech (TTS).

3.1 Speech-to-Text (STT)

We employ Whisper (Radford et al., 2022), a transformer-based STT model trained on a large multilingual corpus. Whisper generates robust transcription along with precise word-level timing annotations, including detailed start and end times. These annotations provide essential temporal cues for preserving the temporal structure of original speech.

3.2 Neural Machine Translation (NMT)

After obtaining transcribed text and timestamps from the STT output, the NMT module translates the source text into the target language. Although traditional NMT systems generate fluent translations through robust contextual understanding, they typically do not synchronize the translation length

Algorithm 1 Duration-based Translation (*DT*)

```
Require: Source text $T_{src}$, Source duration $D_{src}$, Target language $L_{tgt}$

  $P_{pred} \leftarrow$ PredictPhonemes$(D_{src}, L_{tgt})$
  $T_{ref} \leftarrow$ Translate$(T_{src}, L_{tgt})$
  $T_{tgt} \leftarrow T_{ref}$
  $iter \leftarrow 0$
  $d_{tgt} \leftarrow |$CountPhonemes$(T_{tgt}) - P_{pred}|$
  while $d_{tgt} > \delta$ and $iter < MAX$ do
      if CountPhonemes$(T_{tgt}) > P_{pred}$ then
          $C \leftarrow$ Shorter$(T_{tgt}, 3)$
      else
          $C \leftarrow$ Longer$(T_{tgt}, 3)$
      end if
      $F \leftarrow [\,]$
      for each $c$ in $C$ do
          $d_c \leftarrow |$CountPhonemes$(c) - P_{pred}|$
          if $d_c < d_{tgt}$ then
              $F$.append$(c)$
          end if
      end for
      if $F = \emptyset$ then
          break
      end if
      $T_{tgt} \leftarrow \arg\max_{c \in F}$ COMET$(T_{src}, c, T_{ref})$
      $iter \leftarrow iter + 1$
      $d_{tgt} \leftarrow |$CountPhonemes$(T_{tgt}) - P_{pred}|$
  end while
  $T_{aligned} \leftarrow$ AlignPauses$(T_{tgt}, T_{src}, L_{tgt})$
  return $T_{aligned}$
```

with the source speech duration. To address this limitation, we propose a method that explicitly considers both the duration and pause information of the source speech during translation. Algorithm 1 provides the iterative process for generating duration-aligned translations.

Duration-based Translation (*DT*) To accurately estimate actual spoken durations is difficult when relying solely on the number of characters or words. To overcome this, we propose a duration-based length control strategy that estimates the optimal phoneme count from the source speech duration and the typical speaking rate of the TTS system. This ensures that the translated speech closely matches the original duration. First, we analyze the start and end timestamps of each word obtained from the STT module to calculate the total duration of the source sentence, denoted as D_{src}. We then predict the appropriate number of phonemes for the translation by considering both the source speech duration and the average speaking rate of the TTS system in the target language. For instance, if the source segment lasts 2 seconds and the TTS speaking rate is 10 phonemes per second, the translated segment should ideally contain approximately 20 phonemes. To achieve this, we use an iterative translation with an LLM, dynamically refining translations by lengthening or shortening phrases as needed. If the translation is too long or too short, we instruct the LLM to eliminate extraneous phrases or include additional modifiers and explanations. To optimize this process, multiple candidate translations are generated in parallel, and we select the one that best preserves the intended meaning while closely matching the estimated phoneme count, iterating until the translation reaches the desired length. Through this iterative process, we obtain an output that satisfies both temporal alignment with the source and high translation quality.

Pause Integration Although our *DT* achieves accurate duration alignment with the source, we additionally consider natural pause positions to enhance temporal synchronization, since natural speech typically contains pauses for breathing or emphasis. Therefore, we identify pauses in the source speech by analyzing word-level timestamps from the STT module, and then incorporate these pauses into the translation. However, each language has a distinct sentence structure, and directly transferring pause locations from the source can lead to awkward or unnatural segments. To address this, our framework leverages an LLM to determine appropriate pause positions in the translation based on its linguistic structure. If the original speech contains a pause after a certain phrase or clause, the translation is correspondingly segmented at a suitable linguistic boundary to insert an appropriate pause. During TTS synthesis, we insert pauses with precise durations at the identified positions in the translated segments, replicating the rhythm of the original speech and ensuring a natural, synchronized audio output. With pause integration, the synthesized speech naturally preserves the original speaker's pause patterns, resulting in fluent and isochronous dubbing.

Direction	**Speech Overlap**			**COMET**		
	DT	***PT***	***GPT-4o***	***DT***	***PT***	***GPT-4o***
en → es	0.899	0.865	0.774	0.741	0.729	0.784
en → ko	0.898	0.874	0.845	0.838	0.848	0.852
es → en	0.939	0.878	0.698	0.789	0.798	0.833
es → ko	0.933	0.868	0.732	0.860	0.866	0.879
ko → en	0.921	0.905	0.876	0.843	0.846	0.867
ko → es	0.902	0.902	0.820	0.801	0.788	0.841
Avg. → en	**0.930**	0.891	0.787	0.816	0.821	**0.850**
Avg. → es	**0.900**	0.883	0.797	0.771	0.757	**0.813**
Avg. → ko	**0.915**	0.871	0.788	0.849	0.857	**0.865**

Table 1: COMET and Speech Overlap scores for six translation directions under three configurations: Duration-based Translation (*DT*), Phoneme-based Translation (*PT*), and *GPT-4o*. Higher speech overlap indicates tighter temporal alignment, while higher COMET indicates better translation quality.

3.3 Text-to-Speech (TTS)

We use ElevenLabs[3] for speech synthesis. For voice cloning, we use UTMOS (Saeki et al., 2022) quality scores to select appropriate samples within the interquartile range since real-world speech samples often contain significant background noise. Finally, we perform speech-to-speech conversion to restore the original speaker's timbre and prosody, producing naturally synchronized dubbed audio.

4 Experiments

4.1 Settings

Dataset For our experiments, we use the Multilingual Interpretation and Translation Reading-style Dataset provided by AI Hub[4]. This dataset consists of sentence triplets in Korean, English, and Spanish, where each triplet conveys identical semantic content but exhibits natural variations in utterance durations due to differences in linguistic structures and speaking rates. This variability in speech durations provides a suitable testbed for evaluating system performance under temporal constraints. Additionally, the diverse language pairs allow for a comprehensive evaluation of multilingual dubbing capabilities.

From this dataset, we randomly sample 100 triplets for each language. Within each triplet, one language serves as the target, and the other two serve as source languages. Extending this process across multiple triplets ensures a comprehensive evaluation of the model's translation performance across diverse language pairs.

Baselines We compare the following three configurations:

- ***GPT-4o***: Translates directly using GPT-4o without any explicit length constraints, primarily optimizing lexical accuracy and naturalness.
- **Phoneme-based Translation (*PT*)**: Translates iteratively using GPT-4o, explicitly matching the phoneme counts of source and target segments.
- **Duration-based Translation (*DT*)**: Translates iteratively using GPT-4o, dynamically predicting phoneme counts from the source speech duration.

Evaluation Metrics To assess translation quality, we use COMET (Rei et al., 2020), which measures semantic alignment and fluency compared to reference texts. COMET is preferred over BLEU (Papineni et al., 2002) for its superior handling of morphologically rich languages like Korean and cross-lingual evaluation consistency. To measure temporal alignment between the source and the dubbed speech, we adopt speech overlap metric as follows:

$$\text{SO} = 1 - \frac{|\text{source duration} - \text{dub duration}|}{\text{source duration}} \quad (1)$$

This equation computes the absolute difference between the source and dubbed speech durations, normalizes it by the source duration, and inverts the

[3] https://elevenlabs.io/

[4] https://www.aihub.or.kr/

value so that higher scores indicate tighter synchronization. We compute this metric for each sample and report the average across the test set.

4.2 Experimental Results

Table 1 presents the COMET and speech overlap results in six translation directions across English, Spanish, and Korean. When applying *DT*, our approach achieves the highest speech overlap across most translation directions, indicating that *DT* effectively ensures precise temporal alignment. By contrast, *GPT-4o* achieves higher COMET scores on average, but exhibits significantly lower speech overlap, highlighting the inherent trade-off between translation quality and temporal alignment.

Length Control We first analyze the impact of length constraints by comparing the length-constrained approaches, *DT* and *PT*, with the unconstrained *GPT-4o* approach. Without explicit length constraints, COMET scores generally improve as translations have more lexical freedom, but speech overlap substantially decreases. On average across all language pairs, *DT* and *PT* respectively improve speech overlap by approximately 16.15% and 12.14% over *GPT-4o*. This consistent improvements clearly demonstrate that imposing explicit length constraints is essential for achieving precise synchronization and naturalness in automatic dubbing.

Duration-based Phoneme Estimation We then analyze the effect of explicitly predicting and matching the translation's phoneme count based on the source speech duration, compared to directly matching phoneme counts between the source and target segments. Although COMET differences between *DT* and *PT* are minor, the advantage of *DT* becomes evident in terms of speech overlap. On average across all language pairs, *DT* achieves a relative improvement of approximately 3.75% in speech overlap compared to *PT*. For instance, improvements reach up to 7.49% in Spanish-to-Korean and 2.75% in English-to-Korean translations. These results clearly illustrate that matching phoneme counts alone, without considering speech duration, is insufficient for precise temporal alignment. By explicitly estimating phoneme counts from source durations, *DT* consistently improves synchronization while maintaining competitive translation quality, highlighting clear advantages for real-world dubbing.

Method	**Avg. MOS**	**Std. Dev.**
Proprietary System	5.83	1.51
DT	**6.57**	1.46

Table 2: Subjective evaluation results on eight video clips, rated by thirty participants. Scores are reported as mean opinion scores (MOS) with standard deviation; higher scores indicate better quality.

4.3 Human Evaluation

We also measure a Mean Opinion Score (MOS) on eight video clips to compare our method against a proprietary auto-dubbing system. Thirty multilingual participants watch each dubbed clip in randomized order and rate the overall dubbing quality on a ten-point scale, considering both naturalness (e.g., fluency, prosody) and translation accuracy.

Table 2 shows the results. Our approach achieves a higher average MOS compared to the proprietary system, suggesting that explicitly controlling translation length based on source durations and adjusting phoneme counts yields more natural and coherent dubbed speech.

5 Conclusion

In this paper, we propose an end-to-end automatic dubbing framework incorporating *Duration-based Translation (DT)*, a novel translation approach designed to achieve accurate temporal alignment by dynamically adjusting phoneme counts based on source speech durations. Our framework employs a phoneme count predictor that estimates the optimal translation length considering linguistic context and the TTS system's speaking rate. Experiments across multiple language pairs demonstrate that *DT* significantly improves temporal alignment while maintaining competitive translation quality compared to existing methods. These results highlight the practical advantages of *DT* for real-world multilingual dubbing scenarios.

Limitations

Our approach is dependent on the specific speaking rate and phoneme-generation characteristics of the underlying TTS system. Changes or inaccuracies in the TTS engine can therefore directly impact temporal alignment quality, potentially requiring recalibration. In future work, we plan to investigate methods to reduce this dependency and improve robustness across different speech synthesis systems.

References

Frederic Chaume. 2008. Synchronization in dubbing: A translational approach. In *Topics in audiovisual translation*, pages 35–52. John Benjamins Publishing Company.

Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel Melaku Lakew, and Marcello Federico. 2023. Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing. *CoRR*, abs/2302.12979.

Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isometric mt: Neural machine translation for automatic dubbing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6242–6246. IEEE.

Yuang Li, Jiaxin Guo, Min Zhang, Ma Miaomiao, Zhiqiang Rao, Weidong Zhang, Xianghui He, Daimeng Wei, and Hao Yang. 2024. Pause-aware automatic dubbing using LLM and voice cloning. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 12–16, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2019. Prosodic phrase alignment for machine dubbing. *arXiv preprint arXiv:1908.07226*.

Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico. 2023. Improving isochronous machine translation with target factors and auxiliary counters. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 37–41. ISCA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Zhiqiang Rao, Hengchao Shang, Jinlong Yang, Daimeng Wei, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhengzhe Yu, Zhanglin Wu, Yuhao Xie, Bin Wei, Jiawei Zheng, Lizhi Lei, and Hao Yang. 2023. Length-aware NMT and adaptive duration for automatic dubbing. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 138–143, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isochrony-aware neural machine translation for automatic dubbing. In *Interspeech 2022*, pages 1776–1780.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. 2022. Prosodic alignment for off-screen automatic dubbing. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 496–500. ISCA.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13772–13779.