Pause-Aware Automatic Dubbing using LLM and Voice Cloning

Yuang Li, Jiaxin Guo, Min Zhang, Miaomiao Ma, Zhiqiang Rao Weidong Zhang, Xianghui He, Daimeng Wei, Hao Yang Huawei Translation Services Center, China

{liyuang3, guojiaxin1, zhangmin186, mamiaomiao, raozhiqiang, zhangweidong17, hexianghui, weidaimeng, yanghao30}@huawei.com

Abstract

Automatic dubbing aims to translate the speech of a video into another language, ensuring the new speech naturally fits the original video. This paper details Huawei Translation Services Center's (HW-TSC) submission for IWSLT 2024's automatic dubbing task, under an unconstrained setting. Our system's machine translation (MT) component utilizes a Transformerbased MT model and an LLM-based post-editor to produce translations of varying lengths. The text-to-speech (TTS) component employs a VITS-based TTS model and a voice cloning module to emulate the original speaker's vocal timbre. For enhanced dubbing synchrony, we introduce a parsing-informed pause selector. Finally, we rerank multiple results based on lip-sync error distance (LSE-D) and character error rate (CER). Our system achieves LSE-D of 10.75 and 12.19 on subset1 and subset2 of DE-EN test sets respectively, superior to last year's best system.

1 Introduction

The task of automatic dubbing is to translate spoken language in a video into another language such that the translated speech can be seamlessly blended with the original video. A unique aspect of dubbing is isochrony, which refers to the property that the speech translation is time-aligned with the original speaker's visual cues. The spoken words should match the speaker's lip movements, ensuring the audio is heard when the lips move and is silent when they don't.

To address this challenge, a unified model that simultaneously processes translations and speech timing is optimal, allowing for adjustments in translation to fit timing constraints. Chronopoulou et al. (2023) accomplish this by simply binning target phoneme durations and interleaving them with target phonemes during training and inference. Pal et al. (2023) enhance this approach by predicting the durations of phonemes as target factors. However, these methods fail to utilize pre-trained machine translation (MT) models and large language models (LLM) that are trained on massive text corpora. Moreover, constructing large-scale datasets with phoneme duration labels is challenging, thus limiting the translation quality. Therefore, a disentangled approach that considers MT and dubbing synchrony separately can achieve better results. Our system (Rao et al., 2023) from last year first generated a set of translation candidates and later reranked them based on speech overlaps, achieving better mean opinion scores (MOS) than the baseline systems. Therefore, this year we extend last year's system by using more advanced pre-trained models and a more sophisticated pause-aware dubbing pipeline.

Specifically, our method comprises the following key components:

- A Transformer-based MT (Machine Translation) model, which is a fine-tuned version of NLLB-1.3B on the CoVoST2 dataset (Changhan Wang, 2020).
- An LLM-based post-editor that modifies the lengths of translations.
- A VITS-based (Kong et al., 2023) TTS model that is non-autoregressive and supports speed control.
- A voice cloning (VC) module based on Open-Voice (Qin et al., 2023), ensuring that the input speech and output speech share the same tone color.
- A pause-aware dubbing pipeline that identifies potential split points using sentence parsing.
- A reranking method based on LSE-D and CER.

In this paper, we provide detailed analyses of the components mentioned above. Our system



Figure 1: Diagram illustrating the process of automated dubbing: (a) without accounting for pauses; (b) with consideration of pauses. (Note that the ASR results are provided by the organizer in this track.)

achieves an LSE-D of 10.75 on subset1 and 12.19 on subset2 of the DE-EN test sets, respectively, outperforming last year's best system. Additionally, we take into account the vocal timbre of the speech, which can enhance the perceptual quality.

2 Methods

2.1 System Design

Figure 1 (a) shows the naive automatic dubbing system which assumes that the speech of the video does not have obvious pauses. First, an automatic speech recognition (ASR) model transcribes the source speech. The result of this step is provided by the organizer. Then, an MT model translates the source German (DE) text into the target English (EN) text, followed by an LLM that is prompted to change the length of the translation. Utilizing MeloTTS ¹, the target speech is synthesized and its duration is compared with the original speech to determine the speed factor. Finally, we regenerate the target speech, convert the tone color, and fill the audio into the original video based on timestamps from the voice activity detection (VAD) ² system.

Figure 1 (b) illustrates the pause-aware dubbing system. Unlike the naive system, it integrates a pause selector. This selector generates an index of potential word positions that best align with the pauses in the original speech. To avoid unnatural sentence breaks, sentence parsing is employed to determine groups of words that should remain together. Finally, the TTS model is utilized to produce audio clips for each text segment, with the speed factor calculated for each independently. Algorithm 1 Pause Selector

Require: t_{pause} , $text_{MT} = \{w_1, ..., w_n\}$ split = $\{t_1, ..., t_{n-1}\}$, T_{src} , T_{tts} 1: $PP, VP, NP = Parsing(text_{MT})$ 2: index = SplitPoint(PUNC, PP, VP, NP)

3:
$$i = argmin(abs(\frac{t_{pause}}{T_{src}} - \frac{\mathbf{spint}_i}{T_{tts}})) \quad i \in \mathbf{index}$$

4: **return** *i*

2.2 Pause Selector

Algorithm 1 provides the details of the pause selector. Given the time of the pause (t_{pause}) predicted by VAD, the translation $(text_{MT})$, the word-level timestamps (**split**) of synthetic speech predicted by a CTC-based aligner from WhisperX (Bain et al., 2023), and the duration of source and generated speech (T_{src}) and (T_{tts}), we first use sentence parsing ³ to obtain the prepositional phrases (PP), verb phrases (VP), and noun phrases (NP). The possible split index can be only after these phrases and punctuations. Then, we select the best index that minimizes the distance between the normalized word time by duration and the normalized time of the pause.

2.3 LLM-based Post-Editor

You are a professional German-English translator and skilled proofreader. Now you are given the original German text and its English translation. Please improve the translation and make it more **complex/simple** without explaining.

Source (German): "{DE}" Initial Translation (English): "{EN}" Revised Translation (English):

Table 1: Prompt for LLM-based post-editor.

LLM (Touvron et al., 2023; Zeng et al., 2023) is known for its exceptional zero-shot and few-shot capabilities, meaning it can perform downstream tasks using a prompt that describes the task or a few examples. In the context of automatic dubbing, we use LLM to generate translations with different lengths so that we can select the one that results in the best lip-sync accuracy. The input prompt for the LLM is shown in Table 1. We first describe

¹https://github.com/myshell-ai/MeloTTS

²https://github.com/snakers4/silero-vad

³https://github.com/Halvani/ Constituent-Treelib

the task and the role of the LLM as a translator and a proofreader. Then, we instruct it to make the translation more complex or simple. Finally, we provide the source and translated text. We use "complex" and "simple" as indicators of output length as they contribute to better stability than "longer" and "shorter".

2.4 TTS and VC

We use MeloTTS, which is based on the architecture of VITS (Kim et al., 2021; Kong et al., 2023). VITS leverages variational autoencoder, adversarial learning, normalizing flow, and stochastic duration predictor to generate realistic speech in an end-to-end manner without relying on external word alignment and a vocoder. To convert the voice into the desired tone color, we adopt Open-Voice (Qin et al., 2023), which disentangles the tone color information in the encoder. The target speaker embedding is integrated into the decoder.

2.5 Rerank

We use LSE-D and CER to select the final synthetic speech from multiple candidates. The CER is computed between the original ASR transcription and the transcription of the generated speech. For subset1, there are no obvious pauses, so we only use system (a) as shown in Figure 1. For subset2, which contains notable pauses, we use both systems (a) and (b) in Figure 1. For the same translation, we only use system (b) if we do not observe a decline in CER and note an improvement in LSE-D compared to system (a). To rank multiple translations of different lengths, we use the average rank determined by LSE-D and CER and select the translation with the lowest rank. Note that we use CER rather than word error rate to mitigate the influence of the ASR model's limited ability to recognize out-of-vocabulary words. During rerank, we considered four translations: the original translation, the translation using LLM-based post editor without indicating the output length, the "complex" translation, and the "simple" translation.

3 Experimental Setups

We fine-tuned the NLLB-1.3B ⁴ model for 20 epochs on the CoVoST2 (Changhan Wang, 2020) DE-EN subset, using a learning rate of 3×10^{-5} and a batch size of 512. For the LLM-based post-editor,

⁴https://huggingface.co/facebook/

nllb-200-distilled-1.3B

when employing the "complex" indicator, we sampled three answers and selected the one with the highest Comet score (Rei et al., 2020) compared to the original translation. For the "short" indicator, we sampled only once. When adjusting the speech speed, we set the lower bound to $0.75 \times$ and the upper bound to $2.5 \times$. We adopted several evaluation metrics: the BLEU score and the Comet score to evaluate MT quality, and the lip-sync error distance (LSE-D) (Chung and Zisserman, 2017)⁵ and ASR character error rate (CER) to measure dubbing performance. We used the Wav2Vec2base model ⁶, fine-tuned on LibriSpeech, as the ASR model, which utilizes a character-level vocabulary. We opted not to use a more advanced ASR model since the less robust model is more sensitive to speech quality. During rerank, we considered four translations: the original translation, the translation using LLM-based post editor without indicating the output length, the "complex" translation, and the "simple" translation. Additionally, we attempted to enhance the speech by applying denoising and audio super-resolution techniques ⁷, which remove noise and upscale the audio from 16kHz to 44.1kHz.

4 Experimental Results

4.1 Performance of MT and LLM-based Post-Editor

As shown in Table 2, the NLLB-1.3B model, fine-tuned on the target-domain CoVoST2 dataset, achieves high translation quality with BLEU scores of 46.37 and 44.03 on subset1 and subset2, respectively, and Comet scores of 89.29 and 88.01, respectively. When using an LLM to post-process the translations, we observe a decrease in BLEU scores, especially for longer translations. However, we find that the Comet scores are similar to those of the unmodified translations, indicating that the LLM effectively performs paraphrasing without changing the meaning of the translations.

4.2 Results for Pause-Aware Automatic Dubbing

For subset2, we observe that the pause-aware automatic dubbing pipeline (Dubbing (b)) contributes

⁶https://huggingface.co/facebook/ wav2vec2-base-960h

⁵When computing LSE-D, we used the video with subtitles.

⁷https://github.com/resemble-ai/ resemble-enhance

	Subset1				Subset2							
	MT		Dubbing (a)		MT		Dubbing (a)		Dubbing (b)		Dubbing (a + b)	
Method	BLEU↑	Comet↑	LSE-D↓	$CER\downarrow$	BLEU↑	Comet↑	LSE-D↓	$CER\downarrow$	LSE-D↓	$CER\downarrow$	LSE-D↓	$CER\downarrow$
NLLB (fine-tune)	46.37	89.29	10.92	5.68	44.03	88.01	13.88	3.93	12.27	4.47	12.39	3.59
LLM-PE	43.03	89.42	11.03	5.78	40.75	88.00	12.90	3.97	12.25	4.21	12.46	3.51
LLM-PE (simple)	44.50	88.01	10.97	6.35	43.88	87.98	12.89	4.47	12.22	4.61	12.38	3.93
LLM-PE (complex)	19.67	84.08	11.12	4.75	18.88	83.74	13.05	3.60	12.35	4.46	12.73	3.28
Rerank (LSE-D)	41.18	88.15	10.62	5.70	39.13	87.79	/	/	/	/	11.96	3.76
Rerank (CER)	29.42	85.88	11.05	3.76	29.32	85.03	/	/	/	/	12.62	2.36
Rerank (LSE-D&CER)	38.13	87.91	10.75	4.62	38.60	87.13	/	/	/	/	12.19	3.02
+ Enhance	/	/	11.18	5.39	/	/	/	/	/	/	12.52	4.07
- VC	35.10	87.91	10.86	4.08	39.09	87.28	/	/	/	/	12.14	2.73

Table 2: Performance of MT and dubbing measured by BLEU score, Comet score, LSE-D, and ASR-CER (%). Rerank is applied to the results that correspond to the first four rows.

to a significantly lower LSE-D than the naive pipeline (Dubbing (a)). For instance, with pauseaware dubbing, the LSE-D decreases from 13.88 to 12.27 for the original translation. However, there is an increase in CER. The possible reason could be that the pauses in the translation may be unnatural, or the TTS model's ability to generate speech for incomplete sentences is limited. Therefore, we combine the two systems. For the same translation, we only use system (b) if we do not observe a decline in CER and note an improvement in LSE-D compared to system (a). This combination method (Dubbing (a + b)) results in the lowest CERs, and the LSE-D is also notably better than the naive system (a).

4.3 Results for Rerank

LSE-D measures the synchronization of speech with video, while CER assesses speech intelligibility. Employing either metric for reranking could enhance the results according to their respective evaluations. Using their average rank can achieve a balance between them. For subset1 and subset2, the final dubbed videos achieve LSE-D scores of 10.75 and 12.19, respectively, and CERs of 4.62% and 3.02%, respectively. It is worth noting that the CER for longer speeches tends to be lower due to more contextual information, while the LSE-D tends to be higher as it is more difficult to align the pauses.

4.4 Alternative Systems

We carried out ablation studies and provided alternative systems in our submission. When VC is not used, the LSE-D is similar to the complete system. The CER is notably lower because the sole TTS model provides better speech quality, whereas the VC model can introduce some noise. However, without VC, using a female's voice for a male speaker is unreasonable. Our TTS model operates at a sample rate of 16kHz. To improve the subjective listening experience, we adopted an audio super-resolution model to enhance it to 44.1kHz. Perceptually, higher frequencies contribute to better quality. However, we found that audio superresolution negatively impacts the LSE-D and CER, although we do not observe noticeable distortion in the audio samples.

5 Discussion

Compared to last year's system, which utilized a length-aware MT system that employed a length tag to indicate the desired output length, this year's approach aims to enhance translation quality by fine-tuning a pre-trained MT model rather than training one from scratch. Although we attempted to incorporate length tags in the fine-tuning process, we found that they failed to produce translations with varying lengths due to the limited number of epochs and fine-tuning data. Consequently, we used an LLM which has robust rewriting capabilities.

We submitted a single entry for the English-Chinese subtask, which presents significantly greater challenges than the German-English subtask due to factors such as long-form video, speaker changes, and background music. To address these challenges, we enhanced our automatic dubbing system with an open-source diarization model (Desplanques et al., 2020), a source separation tool (Takahashi and Mitsufuji, 2017), and a TTS API ⁸. However, given the complexity of the task and the lack of labeled test set, we have not provided a detailed analysis. In movie dub-

⁸https://github.com/rany2/edge-tts

bing, it is crucial that the emotion of the dubbed speech matches that of the original speech, therefore, expressive TTS models are preferred. We evaluated the Seamless Expressive model (Barrault et al., 2023), however, we observed that the speech quality was inconsistent, and for non-English languages, the speech did not sound native.

6 Conclusion

In this paper, we propose a novel pause-aware automatic dubbing system that ensures translated speech signals are not only accurate but also maintain the timbre of the original speech. The key components involve a novel pause selector, informed by parsing, to align dubbing with the video's pace, a VC model to convert the tone color, and an LLM to provide translation candidates. For future work, we plan to carry out more systematic experiments on long-form, movie-like videos and provide more expressive dubbed videos.

References

- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Proc. InterSpeech*.
- Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.
- Juan Pino Changhan Wang, Anne Wu. 2020. Covost2 and massively multilingual speech-to-text translation.
- Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel M Lakew, and Marcello Federico. 2023. Jointly optimizing translations and speech timing to improve isochrony in automatic dubbing. *arXiv preprint arXiv:2302.12979*.
- Joon Son Chung and Andrew Zisserman. 2017. Out of time: automated lip sync in the wild. In Proc. ACCV Workshop, pages 251–263. Springer.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. ICML*, pages 5530–5540.

- Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. Vits2: Improving quality and efficiency of single-stage textto-speech with adversarial learning and architecture design. arXiv preprint arXiv:2307.16430.
- Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico. 2023. Improving isochronous machine translation with target factors and auxiliary counters. *arXiv preprint arXiv:2305.13204*.
- Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.
- Zhiqiang Rao, Hengchao Shang, Jinlong Yang, Daimeng Wei, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhengzhe Yu, Zhanglin Wu, Yuhao Xie, et al. 2023. Length-aware nmt and adaptive duration for automatic dubbing. In *Proc. IWSLT*, pages 138–143.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.
- Naoya Takahashi and Yuki Mitsufuji. 2017. Multi-scale multi-band densenets for audio source separation. In *Proc. WASPAA*, pages 21–25. IEEE.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. *Proc. ICLR*.