Japanese-to-English Simultaneous Dubbing Prototype

Xiaolin Wang and **Masao Utiyama** and **Eiichiro Sumita**
Advanced Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan
{xiaolin.wang,mutiyama,eiichiro.sumita}@nict.go.jp

Abstract

Live video streaming has become an important form of communication such as virtual conferences. However, for cross-language communication in live video streaming, reading subtitles degrades the viewing experience. To address this problem, our simultaneous dubbing prototype translates and replaces the original speech of a live video stream in a simultaneous manner. Tests on a collection of 90 public videos show that our system achieves a low average latency of 11.90 seconds for smooth playback. Our method is general and can be extended to other language pairs.

1 Introduction

Live video streaming over the Internet has become a very important form of communication in human society. It has many advantages such as fast, not constrained by distance, economical and safe.

If the language barrier (Ahmad Abuarqoub, 2019) can be broken down in live video streaming, it will greatly promote global communication. However, the current common solution to cross-language live video streaming is to use automatic simultaneous interpretation (Müller et al., 2016; Wang et al., 2016; Franceschini et al., 2020; Bojar et al., 2021) to display translated subtitles. Reading subtitles at the bottom of the screen is uncomfortable and degrades the viewing experience (Wissmath et al., 2009).

Our simultaneous dubbing prototype aims to help live video streaming break down language barriers. Our prototype translates and replaces the original speech of a live video stream, creating a seamless viewing experience in the target language. Table 1 summarizes what our system is. Our system consists of a complete simultaneous interpretation system and a simplified automatic language dubbing system (Furukawa et al., 2016; Yang et al., 2020; Öktem et al., 2019; Federico et al., 2020). By combining these two technologies, it gains a novel ability of live video streaming in a target language.

Feature	SI	LD	Ours
Speech Recognition	√	√	√
Machine Translation	√	√	√
Low Latency	√		√
Text-to-Speech		√	√
Duration Match		√	√
Audio Rendering		√	
Lip Sync		√	
Live Streaming			√

Table 1: Comparison of automatic simultaneous interpretation (SI), automatic language dubbing (LD) and our system.

Tests on a collection of 90 public videos show that the live streaming from our system achieves a low average latency of 11.90 seconds and meets a smoothness criterion. Therefore, our system can be widely used in fields such as news broadcasting, conferences and education. Furthermore, our method is general and can extend to other language pairs.

The main contributions of our work include,

- implementing a first simultaneous dubbing prototype for multi-language live video streaming;
- developing evaluation metrics for the latency, smoothness and duration matching of simultaneous dubbing;
- proposing an adaptive playback method to balance latency and smoothness.

The rest of this paper is organized as follows. First, Section 2 reviews related works. Then, Section 3 describes our method for implementing simultaneous dubbing. After that, Section 4 tests our system on a collection of 90 public videos in

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 169–178
July 10-12, 2023 ©2023 Association for Computational Linguistics

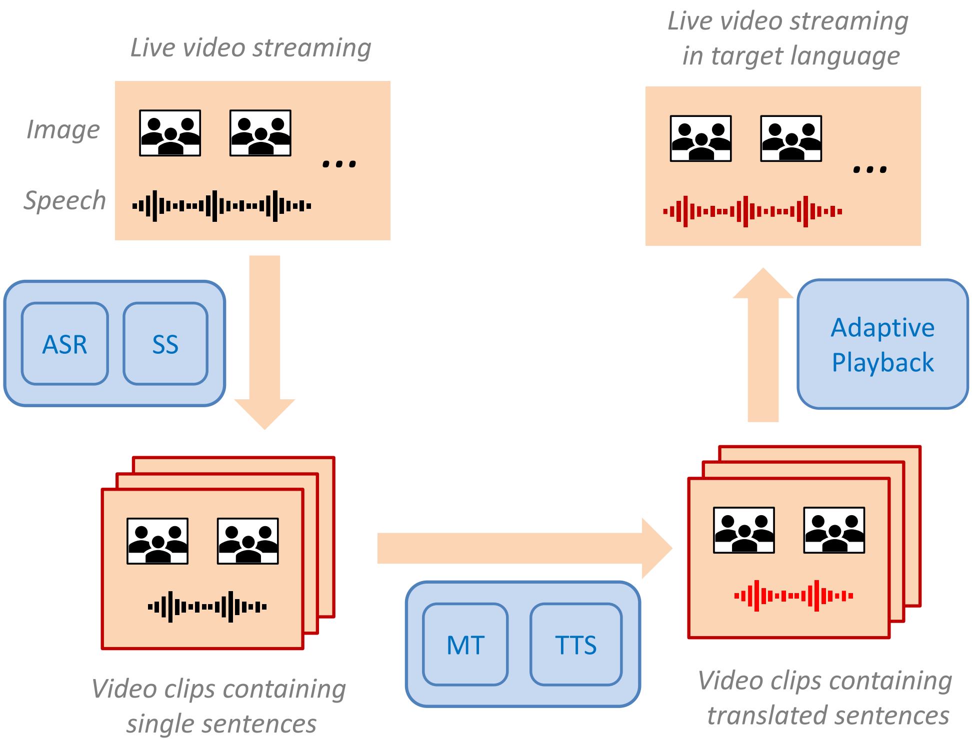


Figure 1: Implementation of simultaneous dubbing using automatic speech recognition (ASR), sentence segmentation (SS), machine translation (MT), text-to-speech (TTS) and adaptive playback.

terms of latency, smoothness and duration matching. Finally, Section 5 concludes this paper with a description on future works.

2 Related Works

Automatic simultaneous interpretation and automatic language dubbing are the two topics most closely related to our work.

2.1 Automatic Simultaneous Interpretation

Simultaneous interpretation is a hot topic. Due to space limitations, we only review some selected practical systems.

Professor Alex Waibel from the Karlsruhe Institute of Technology (KIT) demonstrates a simultaneous interpretation system that automatically translates lectures from German to English in 2012 (Figure 2a) [1]. The transcripts are shown on the left part of the window and the translation is shown below.

Microsoft Meetings pilots live translated subtitles in 2022 (Figure 2b) [2]. With this new feature, users can select a translation language for live subtitles. This feature helps users fully participate in meetings where the spoken language may not be their most comfortable language to use. Google Meet has a similar feature [3].

Wang et al. (2022) demonstrate a multimodal simultaneous interpretation system that annotates translation with speakers (Figure 2c). Due to the delays in the process of simultaneous interpretation, it is sometimes difficult for users to trace the translation back to speakers. Thus, the system explicitly presents "who said what" to users.

Our work differs from these related works by presenting translation as dubbing, whereas related works present translation as subtitles. We believe our method can be incorporated into these related works to bring better services to users.

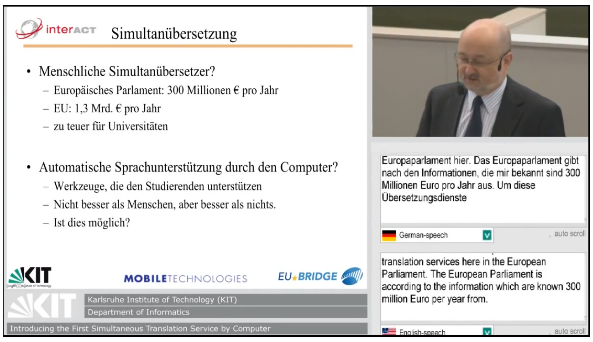
2.2 Automatic Language Dubbing

Automatic Language Dubbing commonly operates on entire video (Yang et al., 2020; Öktem et al., 2019; Federico et al., 2020) whereas our work operates on video streams and generates output in low latency. In addition, due to the complexity of the task, manually correction and adjustment are

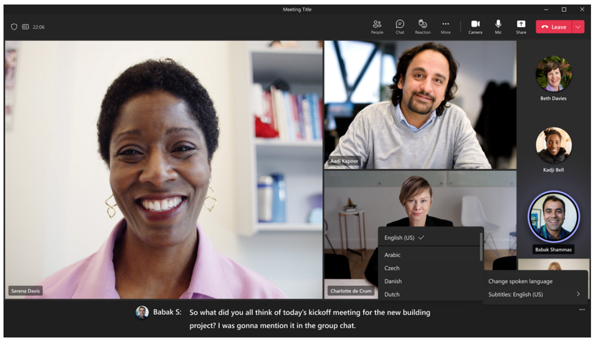
[1] https://www.youtube.com/watch?v=GHeHiPh3u0s

[2] https://techcommunity.microsoft.com/t5/microsoft-teams-public-preview/now-in-public-preview-live-translated-captions-in-meetings/m-p/3620055

[3] https://workspaceupdates.googleblog.com/2022/01/live-translated-captions-in-google-meet-generally-available.html

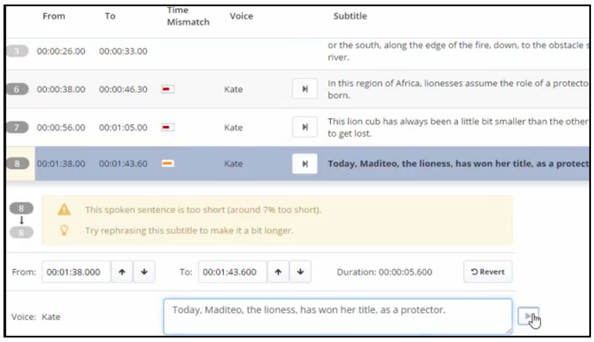


(a) KIT SI of lectures



(b) Microsoft SI of meetings

(c) SI with multimodal speaker recognition



(d) VideoDubber LD

Figure 2: Automatic simultaneous interpretation (SI) and language dubbing (LD) systems.

often required, such as VideoDubber (Figure 2d) [4], whereas our work is fully automatic.

3 Methods

Our prototype accomplishes simultaneous dubbing through three main steps as (Figure 1),

1. Segmenting the source video stream into video clips that contain one single sentence using automatic speech recognition (Hinton et al., 2012; Graves and Jaitly, 2014) and sentence segmentation (Sridhar et al., 2013; Iranzo-Sánchez et al., 2020). For automatic speech recognition, we use the Transformer-based (Vaswani et al., 2017) acoustic model and the seq2seq criterion (Sutskever et al., 2014; Synnaeve et al., 2019) implemented in Flashlight (Pratap et al., 2019)[5]. For sentence segmentation, we replace the backbone network of CytonNSS (Wang et al., 2019)[6] with Transformer to improve accuracy.

2. Generating a translated speech waveform for each sentence using machine translation (Bahdanau et al., 2014; Stahlberg, 2020) and text-to-speech (Wang et al., 2017; Ren et al., 2019). For machine translation, we use the Transformer model implemented in OpenNMT (Klein et al., 2017) [7]. For text-to-speech, we modify the official implementation of VITS (Kim et al., 2021) [8] to generate speech waveforms from speaker embeddings to match the original voice, similar to (Jia et al., 2018).

3. Playing the images and the translated speech waveforms using an adaptive playback method.

The main challenge of simultaneous dubbing is that the output of sentence segmentation (Step 1) and machine translation (Step 2) is irregular in time, but video streaming is constantly consuming data. For example, in the source stream, someone speaks a sentence for about 15 seconds. The system then spends another 5 seconds generating the translated speech waveform. This results in a 20-second data gap in the output stream.

The adaptive playback method addresses this challenge while maintaining low latency (Figure 3).

[4] https://app.videodubber.com/?source=hp_dub_it_now
[5] https://github.com/flashlight/flashlight/tree/main/flashlight/app/asr
[6] https://github.com/arthurxlw/cytonNss
[7] https://github.com/OpenNMT/OpenNMT-py
[8] https://github.com/jaywalnut310/vits

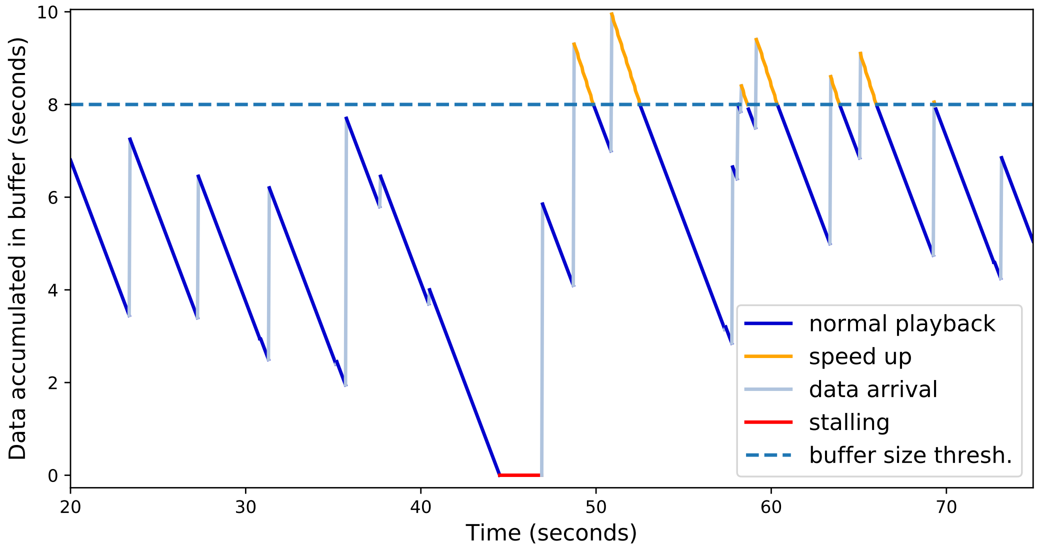


Figure 3: Behaviors of adaptive playback method.

The speed of the playback changes according to the size of the data accumulated in the playback buffer, formulated as,

$$\text{speed} = \begin{cases} 1.0 & \text{if } x < \theta, \\ \alpha & \text{if } x \geq \theta, \end{cases} \tag{1}$$

where x is the amount of data accumulated in the playback buffer. The playback acceleration $\alpha \geq 1$ and the buffer size threshold θ are parameters that control latency and smoothness (Section 4.2).

4 Evaluation

Our system is tested on a collection of 90 public videos of Japanese interviews, speeches, presentations and lectures. The total running time of the collection is approximately 21 hours 45 minutes. The tests are run on a desktop computer equipped with one Intel Xeon E5-2630 V3 CPU and two Nvidia Quadro RTX 4000 GPUs.

The test results are shown in Table 2. The performance of our system is evaluated in terms of latency (Section 4.1), smoothness (Section 4.2) and duration matching (Section 4.3).

Our system presets three modes, Fast, Balance and Quality, for different trade-offs of speed and quality. Users can select the mode according to the application. Table 3 lists the parameters for each mode. Table 7 shows grid search for the buffer size threshold and playback acceleration for the fast mode.

4.1 Latency

Latency is the delay between the input video stream and the output video stream. It is calculated by comparing the start time of each source sentence in the input stream with that of the corresponding translation, formulated as,

$$\text{Latency} = \frac{\sum_{i=1}^{N_{\text{sent}}} T_{i,s} - T_{i,o}}{N_{\text{sent}}}, \tag{2}$$

where N_{sent} is number of the sentences, $T_{i,o}$ and $T_{i,s}$ are the start times of original waveform and synthesized translated waveform, respectively. Table 5 gives an example with a latency of 9.8.

The fast mode on our system achieves an average latency of 11.90 seconds (Table 2). This is relatively fast as the maximum duration of sentences in each video averages 10.76 seconds and the maximum delay of the generated translated speech averages 15.59 seconds on the whole dataset (Table 4). It is difficult to reduce the latency too much below this value while maintaining smooth video streaming.

4.2 Smoothness

The smoothness of the output stream is measured by,

- **# Stall** : the average number of stalls per minute.
- **S. Dur.** : the total duration of stalls per minute.

This follows the researches on assessing the quality of Internet video streaming (Pastrana-Vidal et al., 2004; Qi and Dai, 2006; Moorthy et al., 2012; Seufert et al., 2014; Garcia et al., 2014; Bampis et al., 2017; Zhou et al., 2022)

Mode	Latency	Smoothness		Duration Match		
	(s)↓	# Stall.↓	S. Dur.(s)↓	Fit (%) ↑	D. Fit (%)↑	D. Ex.(%) ↓
Fast	**11.90**	1.21	2.55	89.32	71.43	154.03
Balance	12.90	0.71	1.71	90.60	75.50	146.67
Quality	14.12	**0.49**	**1.38**	**91.50**	**78.43**	**126.16**

Table 2: Evaluation Results.↓ the smaller the better. ↑ the higher the better. (s) seconds.

Mode	Playback		MT
	Buf.(s)	Acc.	# Models
Fast	5.0	x 1.06	1
Balance	7.0	x 1.04	2
Quality	9.0	x 1.02	3

Table 3: Paramteres

Video	Max Dur.(s)	Max Delay(s)
1	13.45	15.05
2	10.66	16.80
3	9.97	16.40
4	11.81	17.40
	...	
87	9.09	14.20
88	8.88	14.45
89	8.81	13.00
90	10.86	18.05
Average	10.76	15.59

Table 4: Maximum duration and processing delay per sentence for each video stream using one machine translation model.

Users tend to tolerate up to three short one-second stalls, or one long three-second stall according to the crowdsourcing-based studies (Hoßfeld et al., 2011). The fast mode of our system is slightly better than this guideline, while the balance mode and the quality mode are well above this guideline (Table 2).

The smoothness of the streaming is influenced by the buffer size threshold and the acceleration in the adaptive playback module. We perform grid search for these two parameters for the fast mode, balance and quality mode, respectively. Table 7 shows the search result for the fast mode. To speed up the search, we record the ready time of each sentence and simulate on the playback module.

4.3 Duration Matching

Language dubbing requires that the duration of each translated speech waveform matches the duration of its source sentence. The duration matching is measured as,

- **Fit** $(\%)$: the percentage of the translated speech waveforms that **fit** in their original durations, formulated as,

$$\frac{N_{\text{Fit}}}{N_{\text{Fit}} + N_{\text{Exceed}}} \times 100\%, \tag{3}$$

where N_{Fit} and N_{Exceed} is the number of translated speech waveforms that fit and exceed the original durations, respectively.

- **D. Fit** $(\%)$: the average percentage of the **durations** for the translated waveforms that **fit** the original durations, formulated as,

$$\sum_{i=1}^{N_{\text{Fit}}} \frac{D_{i,s}}{D_{i,o}} \times 100\%, \tag{4}$$

where $D_{i,s} \leq D_{i,o}$, and they are the durations of synthesized waveforms and original waveforms, respectively.

- **D. Ex.** $(\%)$: the average percentage of the **durations** for the synthesized waveforms that **exceed** the original durations, formulated as,

$$\sum_{j=1}^{N_{\text{Exceed}}} \frac{D_{j,s}}{D_{j,o}} \times 100\%, \tag{5}$$

where $D_{j,s} \geq D_{j,o}$.

Table 6 shows an example of measuring duration matching.

Our system meets the requirement by trying multiple translation candidates for each source sentence. In the fast mode, our system uses the best three candidates that are generated by a machine translation model. In the quality mode, our system employs three machine translation models, that is, nine translation candidates. Table 2 shows that by increasing the number of translation models, the Fit and D. Fit percentages increase and D. Ex. decreases percentage accordingly.

No.	Source Sentence	Translation	Time (s)		
			Start.	Play.	Delay
1	大学教育入門第九章アカデミックプレゼンテーション	Introduction to University Education Chapter 9: Academic Presentation	1.8	11.1	9.3
2	パートフォーの講義になります	Part Four.	6.0	15.3	9.3
3	この講義ではプレゼンテーションの話し方についてまず説明します	In this lecture, we'll start with a presentation.	9.2	18.5	9.3
4	まず事前練習は必ずしましょう	Be sure to do the pre-practice first.	15.3	24.7	9.4
5	お部屋で一人ででもいいのでまずしゃべってみることが大事です	You can do it alone in the room, so it's important to talk to them first.	18.3	29.8	11.5
Average					9.8

Table 5: Example of measuring latency. **Start** time and **Playback** time are measured at the beginning of sentences and translations, respectively.

No.	Source Sentence	Translation	Duration(s)		Dur. Match. (%)		
			Sour.	Trans.	Fit	D.Fit	D.Ex.
1	一般契約ができたのも毎回毎回七社とプレゼン合うんですよね	I was able to make a general contract, and each time I made a presentation with seven companies, right?	4.97	4.18	Yes	84.1	
2	スピードデートみたいな形で三十分から一時間ずつ会っていくんですよ	We meet for thirty minutes to an hour each time in the form of a speed date.	3.45	3.01	Yes	87.2	
3	そのときに僕は世界的な著者になる準備をしてきたし	That's when I was preparing to become a world-class author.	3.70	3.02	Yes	98.3	
4	日本でも実績もあるしほぼいけるんじゃないかなと思うと	I also have a track record in Japan, so I think I'll be almost able to do it.	3.05	3.34	No		109.5
5	もちろん確信は百%あるわけじゃないけど僕はその仲間も助けてくれることもあるし	Of course, I'm not 100 percent sure, but sometimes my friends can also help me.	5.43	3.46	Yes	63.7	
Average					80.0	79.1	109.5

Table 6: Example of measuring duration matching.

	Buffer size threshold (seconds)										
	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
a. Latency (seconds)											
x1.00	14.66	14.66	14.66	14.66	14.66	14.66	14.66	14.66	14.66	14.66	14.66
x1.01	13.22	13.22	13.23	13.26	13.32	13.41	13.54	13.72	13.92	14.13	14.30
x1.02	12.59	12.59	12.61	12.66	12.74	12.87	13.05	13.28	13.54	13.83	14.08
x1.03	12.19	12.20	12.23	12.28	12.39	12.54	12.75	13.00	13.31	13.64	13.94
x1.04	11.90	11.91	11.95	12.02	12.13	12.31	12.53	12.82	13.14	13.50	13.83
x1.05	11.69	11.70	11.73	11.81	11.94	12.13	12.37	12.67	13.02	13.39	13.74
x1.06	11.50	11.52	11.56	11.64	11.78	**11.98**	12.24	12.55	12.92	13.30	13.67
x1.07	11.35	11.37	11.41	11.50	11.65	11.86	12.13	12.46	12.83	13.23	13.61
b. # stalls (per minute)											
x1.00	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
x1.01	0.66	0.68	0.66	0.64	0.62	0.58	0.53	0.51	0.47	0.45	0.43
x1.02	0.98	1.00	0.95	0.91	0.84	0.74	0.66	0.58	0.52	0.48	0.45
x1.03	1.28	1.28	1.22	1.14	1.02	0.90	0.77	0.65	0.56	0.50	0.46
x1.04	1.54	1.53	1.45	1.34	1.19	1.03	0.87	0.72	0.61	0.52	0.48
x1.05	1.83	1.80	1.66	1.52	1.34	1.14	0.94	0.78	0.64	0.53	0.49
x1.06	2.11	2.08	1.91	1.70	1.49	**1.24**	1.02	0.82	0.67	0.56	0.50
x1.07	2.41	2.32	2.12	1.87	1.60	1.34	1.08	0.87	0.69	0.57	0.50
c. Total duration of stalls (seconds per minute)											
x1.00	1.27	1.27	1.27	1.27	1.27	1.27	1.27	1.27	1.27	1.27	1.27
x1.01	1.69	1.67	1.66	1.62	1.58	1.52	1.46	1.40	1.35	1.32	1.30
x1.02	2.20	2.17	2.11	2.03	1.91	1.78	1.65	1.53	1.43	1.36	1.32
x1.03	2.72	2.66	2.56	2.40	2.22	2.01	1.81	1.64	1.50	1.40	1.34
x1.04	3.24	3.14	2.98	2.76	2.50	2.22	1.96	1.73	1.56	1.43	1.36
x1.05	3.76	3.61	3.39	3.10	2.76	2.42	2.10	1.82	1.61	1.47	1.38
x1.06	4.26	4.07	3.78	3.42	3.01	**2.60**	2.22	1.91	1.66	1.49	1.39
x1.07	4.76	4.51	4.16	3.72	3.25	2.77	2.34	1.98	1.71	1.52	1.41

Table 7: Grid search for the optimal buffer size threshold (0.0 - 10.0 seconds) and playback acceleration (x1.00 - x1.07) for the fast mode. The criteria are: **a.** Latency is as small as possible. **b.** # stalls ≤ 3 times per minute. **c.** Total duration of stalls ≤ 3 seconds per minute.

Our system chooses the longest translated speech waveform within the original duration among the candidates. If all the waveforms exceed the original duration, our system will choose the shortest one and truncate its excess to avoid overlapping with the next sentence. Our system does not adjust speech rate as it makes the sound weird and degrades viewing experience.

We have tried controlling the output length of machine translation, similar to (Lakew et al., 2019), but for our Japanese-English language pair, the translation quality drops a lot. We think the reason is that these two languages are so different that the translation cannot be enforced to have a similar length with the source sentence.

5 Conclusion

This paper presents our Japanese-to-English simultaneous dubbing prototype. The system enables low-latency and smooth live video streaming in the target language. We believe this technology will find widespread use in global communications.

In the future, we plan to add optical character recognition to our system. Video streaming often displays some text, such as the slides that appear in a lecture. Text in video streaming is an important source of information for viewers. Therefore, we hope that by recognizing and translating the text in video streaming, our system can provide users with a complete viewing experience in the target language.

Acknowledgements

A part of this work was conducted under the commissioned research program "Research and Development of Advanced Multilingual Translation Technology" in the "R&D Project for Information and Communications Technology (JPMI00316)" of the Ministry of Internal Affairs and Communications (MIC), Japan.

Ethical Considerations

Our system differs from generating deepfake video contents. Viewers can distinguish the dubbed video streams from original video streams, so it is unlikely for others to use our system in harmful ways. The purpose of our system is to deliver information to viewers in their native language, not to generate realistic videos. We do not synchronize lip with speech or render speech with background noise because they would not help with that goal but introduce additional latency in the output. From these two aspects, viewers can tell the dubbed streams from original video streams. Additionally, we place visible annotations on the output stream indicating that it is dubbed by automatic machine translation.

References

I Ahmad Abuarqoub. 2019. Language barriers to effective communication. *Utopía y Praxis Latinoamericana*, 24.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations.*, pages 1–15.

Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. 2017. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. Association for Computational Linguistics.

Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020. From speech-to-speech translation to automatic dubbing. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 257–264, Online. Association for Computational Linguistics.

Dario Franceschini, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, Barry Haddow, Philip Williams, Rico Sennrich, Ondřej Bojar, Sangeet Sagar, Dominik Macháček, and Otakar Smrž. 2020. Removing European language barriers with innovative machine translation technology. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 44–49, Marseille, France. European Language Resources Association.

Shoichi Furukawa, Takuya Kato, Pavel Savkin, and Shigeo Morishima. 2016. Video reshuffling: automatic video dubbing without prior knowledge. In *ACM SIGGRAPH 2016 Posters*, pages 1–2.

M-N Garcia, Francesca De Simone, Samira Tavakoli, Nicolas Staelens, Sebastian Egger, Kjell Brunnström, and Alexander Raake. 2014. Quality of experience and http adaptive streaming: A review of subjective studies. In *2014 sixth international workshop on quality of multimedia experience (qomex)*, pages 141–146. IEEE.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. 2011. Quantification of youtube qoe via crowdsourcing. In *2011 IEEE International Symposium on Multimedia*, pages 494–499. IEEE.

Javier Iranzo-Sánchez, Adria Giménez Pastor, Joan Albert Silvestre-Cerda, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020. Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599–2611.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *CoRR*, abs/2106.06103.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *IWSLT 2019 International Workshop on Spoken Language Translation*.

Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. 2012. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):652–671.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Ricardo R Pastrana-Vidal, Jean Charles Gicquel, Catherine Colomes, and Hocine Cherifi. 2004. Sporadic frame dropping impact on quality perception. In *Human Vision and Electronic Imaging IX*, volume 5292, pages 182–193. SPIE.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE.

Yining Qi and Mingyuan Dai. 2006. The effect of frame freezing and frame skipping on video quality. In *2006 international conference on intelligent information hiding and multimedia*, pages 423–426. IEEE.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2014. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys & Tutorials*, 17(1):469–492.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xiaolin Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2016. A prototype automatic simultaneous interpretation system. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 30–34.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 1–11.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2022. A multimodal simultaneous interpretation prototype: Who said what. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 132–143, Orlando, USA. Association for Machine Translation in the Americas.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Bartholomäus Wissmath, David Weibel, and Rudolf Groner. 2009. Dubbing or subtitling? effects on spatial presence, transportation, flow, and enjoyment. *Journal of Media Psychology*, 21(3):114–125.

Yi Yang, Brendan Shillingford, Yannis M. Assael, Miaosen Wang, Wendi Liu, Yutian Chen, Yu Zhang, Eren Sezener, Luis C. Cobo, Misha Denil, Yusuf Aytar, and Nando de Freitas. 2020. Large-scale multilingual audio visual dubbing. *CoRR*, abs/2011.03530.

Wei Zhou, Xiongkuo Min, Hong Li, and Qiuping Jiang. 2022. A brief survey on adaptive video streaming quality assessment. *Journal of Visual Communication and Image Representation*, page 103526.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2019. Prosodic Phrase Alignment for Machine Dubbing. In *Proc. Interspeech 2019*, pages 4215–4219.