MELD-ST: An Emotion-aware Speech Translation Dataset

Sirou Chen[1†‡] **Sakiko Yahata**[2†] **Shuichiro Shimizu**[2†]
Zhengdong Yang[2] **Yihang Li**[3‡] **Chenhui Chu**[2] **Sadao Kurohashi**[2,4]

[1]Technical University of Munich, Germany [2]Kyoto University, Japan
[3]SenseTime, Japan [4]National Institute of Informatics, Japan

ge23zuh@mytum.de {yahata,sshimizu}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Emotion plays a crucial role in human conversation. This paper underscores the significance of considering emotion in speech translation. We present the MELD-ST dataset for the emotion-aware speech translation task, comprising English-to-Japanese and English-to-German language pairs. Each language pair includes about $10,000$ utterances annotated with emotion labels from the MELD dataset. Baseline experiments using the SeamlessM4T model on the dataset indicate that fine-tuning with emotion labels can enhance translation performance in some settings, highlighting the need for further research in emotion-aware speech translation systems.

1 Introduction

Human conversation naturally involves emotion. An addressee relies on the speaker's multimodal cues, such as vocal tones and facial expressions, to understand the meaning of an utterance. Handling emotions in machine learning systems is therefore considered an important task, as exemplified by NLP tasks such as sentiment analysis and emotion recognition in conversation (Fu et al., 2023).

Considering emotion in translation is also important. For example, the phrase "Oh my God!" can express a wide range of emotions or reactions, including surprise, shock, awe, excitement, distress, etc., to convey a strong emotional response to a situation, whether positive or negative. Because the literal translation of this wouldn't make sense in a different culture, such emotional phrases need to be translated differently depending on the emotion. For instance, the Japanese translation of the phrase showing surprise could be "マジか！ (*majika*)," whereas it could be "やった！ (*yatta*)" when it shows excitement.

Emotion has been studied in machine translation (or text-to-text translation, T2TT) studies (Troiano et al., 2020). However, there has been little focus on emotion in speech translation (ST). ST is a task of translating from speech to text (speech-to-text translation, S2TT) or speech (speech-to-speech translation, S2ST). ST performance has greatly improved over the recent years with significant efforts on datasets (detailed in Section 2) and models (Seamless Communication et al., 2023a,b; Rubenstein et al., 2023; Radford et al., 2022). Although a recent study by Seamless Communication et al. (2023b) focuses on emotion, further community effort is required in this domain.

To address this gap, we present the MELD-ST dataset, which consists of about $10,000$ utterances in English-to-Japanese (En-Ja) and English-to-German (En-De) language pairs, respectively. We extract audio and subtitles from the TV series *Friends*, with emotion labels for each utterance obtained from the MELD dataset (Poria et al., 2019).

We conduct baseline S2TT and S2ST experiments using the SeamlessM4T v2 model (Seamless Communication et al., 2023b). We show that fine-tuning improves the translation performance, and using the emotion labels can enhance the performance in some settings.

2 Related Work

ST performance has significantly improved in recent years owing to the development of datasets, including S2TT datasets such as MuST-C (Di Gangi et al., 2019) or CoVoST 2 (Wang et al., 2021) datasets, as well as S2ST datasets such as CVSS (Jia et al., 2022) or GigaS2S (Chen et al., 2021; Ye et al., 2023; Agarwal et al., 2023) datasets, inter alia. There are also ST datasets focusing on specific aspects of ST, such as gender (Bentivogli et al., 2020) or dialects (Anastasopoulos et al., 2022), as well as specific settings

[†]Equal contribution.
[‡]Work done while at Kyoto University.

August 11-16, 2024 ©2024 Association for Computational Linguistics

	# utts	En speech (h)	Target speech (h)
		En-Ja	
Train	8,069	6.4	6.1
Dev.	1,008	0.8	0.5
Test	1,008	0.7	0.8
		En-De	
Train	9,314	6.9	7.1
Dev.	1,164	0.8	0.9
Test	1,164	0.8	1.0

Table 1: Statistics of the MELD-ST dataset.

such as subtitles (Karakanta et al., 2020) or cross-language dialogue (Shimizu et al., 2023), inter alia.

A recent study by Seamless Communication et al. (2023b) investigates emotion in ST, presenting the SEAMLESSEXPRESSIVE model that captures prosody and preserves vocal style. They created mExpresso and mDRAL corpora as extensions of existing datasets (Nguyen et al., 2023; Ward et al., 2023), as well as automatically aligned and synthetic corpora.

The MELD-ST dataset is based on the MELD dataset (Poria et al., 2019), an emotion recognition dataset of multimodal multi-party conversation based on the TV series *Friends*. It contains videos with English speech and is annotated with English text, sentiment[1] and emotion[2] labels, speaker information, and timestamps based on audio for each utterance.

The key differences between the MELD-ST dataset and existing expressive ST datasets include: 1) Inclusion of emotion labels for each utterance, which can be useful for experiments and analyses. 2) Origin from a TV series in an emotionally rich environment, with translations and acted speech by professionals, making it suitable for a pilot study of emotion-aware ST research. 3) Coverage of the En-Ja language pair, introducing unique challenges such as the need for translation content adjustments, as described in Section 1.

3 MELD-ST Dataset

The MELD-ST dataset is constructed from translations obtained from a Blu-ray disk and emotion labels from the MELD dataset. This section describes the construction process. The dataset statistics are summarized in Table 1.

3.1 Subtitles and Timestamp Extraction

First, we extracted Japanese and German subtitles along with the timestamps indicating when they are displayed. We used off-the-shelf software to obtain them from a Blu-ray disk. The timestamps were directly obtained from the files. Because the subtitles were included as images representing the subtitle text, we used optical character recognition (OCR) tools[3] to extract them.

3.2 Text Cleaning

We cleaned the extracted subtitles by applying some heuristics. Specifically, we excluded speaker names at the beginning of utterances, duplicated subtitles, and apparent OCR errors.

3.3 Alignment with MELD using Timestamps

The MELD dataset contains utterances along with their timestamps. To align MELD utterances with the subtitles extracted above, we first roughly extracted the audio and further processed them for better alignment. Specifically, we followed the following steps: 1) Find utterance candidates (i.e., utterances where there are overlaps between MELD and subtitles timestamps). 2) Extract the audio of the candidates using the timestamps of the subtitles. 3) If there are multiple utterances in the time span, apply CTC segmentation (Kürzinger et al., 2020)[4] on the candidate audio to correct timestamps, and select the candidate with the longest time overlap. More analysis on this process is provided in Appendix A.

3.4 Data Split

We split the obtained utterances to train, development, and test sets. For part of the development and test sets, further manual cleaning was applied.[5] For Japanese data, the contents of the audio and the subtitles were sometimes different, due to the gap between the written and spoken style of the language. We used Whisper (Radford et al.,

[1] negative, neutral, and positive

[2] anger, disgust, fear, sadness, joy, surprise, and neutral

[3] Ja: https://github.com/hrishikeshrt/google_drive_ocr, De: https://github.com/SubtitleEdit/subtitleedit

[4] We used ESPnet (Watanabe et al., 2018) with models espnet/kamo-naoyuki_wsj and espnet/german_commonvoice_blstm.

[5] We manually analyzed non-neutral utterances (i.e., utterances with emotion labels that are not "neutral") in detail. Therefore, most non-neutral utterances are manually checked and corrected, whereas neutral utterances are not.

		Anger	Disgust	Fear	Sadness	Joy	Surprise	Neutral
En-Ja	Train	12.18%	2.95%	2.59%	7.47%	15.91%	11.35%	47.54%
	Dev.	11.81%	2.18%	3.27%	8.23%	17.46%	9.50%	47.52%
	Test	8.43%	3.87%	2.48%	7.24%	18.45%	12.00%	47.52%
En-De	Train	11.76%	2.80%	2.49%	7.04%	16.87%	11.77%	47.26%
	Dev.	10.91%	2.15%	2.66%	8.51%	17.35%	11.17%	47.25%
	Test	8.76%	3.35%	2.75%	7.90%	24.14%	11.25%	47.25%

Table 2: Emotion distribution of our MELD-ST dataset.

2022) to transcribe the audio and manually corrected the errors for part of the development and test sets. For the training set, the subtitles are used despite the style difference.

Emotion label distribution was carefully considered during the data-splitting process, with details provided in Table 2. Almost half of the sentences' emotions are neutral. In the rest, some of the emotion labels are more prevalent than others, like anger, joy, and surprise.

4 Experimental Settings

We conducted S2TT and S2ST experiments using SeamlssM4T. This section provides the details of the experiments and evaluation settings.

4.1 Models for Comparison

Our models were based on SeamlessM4T (Seamless Communication et al., 2023a,b), which supports both S2TT and S2ST tasks. It integrates a massively multilingual T2TT model, an unsupervised speech representation learning model, a text-to-unit encoder and decoder, and a speech resynthesis vocoder. Its different components can be jointly optimized, effectively addressing issues related to cascaded error propagation and domain mismatch. The SeamlessM4T v2 model serves as a test bed for fine-tuning and analysis, and its speech-to-speech translation covers translation from English into 35 languages, including Japanese and German. Because the focus of this paper is to present the MELD-ST dataset with a reasonable baseline, we conducted experiments with the medium model.[6]

We compare the following three settings:

- **No fine-tuning**: We evaluated the SeamlessM4T v2 medium model on the test set of the MELD-ST dataset.
- **Fine-tuning without emotion labels**: We fine-tuned the SeamlessM4T v2 medium model on the MELD-ST dataset without utilizing the emotion labels.
- **Fine-tuning with emotion labels**: We used the emotion labels annotated for each utterance in the original MELD dataset. Following the method of Gaido et al. (2020), a study that investigated the usage of gender information for speech translation, we prepended the gold emotion labels at the beginning of the decoder input sequence during training.[7] During testing, the emotion labels were predicted along with the translations.

The fine-tuning settings included a batch size of 4, evaluation steps of $1,000$, and a maximum of 200 epochs. Fine-tuning was conducted on a single Nvidia A100 80GB GPU. The training process stopped when the loss didn't improve for 10 epochs, and the best checkpoint based on the translation quality of the development set would be used. Each fine-tuning process lasted about 2 hours on one GPU.

The model was fine-tuned using three data settings: En-Ja, En-De, and a mixed dataset combining En-Ja and En-De of the MELD-ST dataset. We fine-tuned the model multiple times with the same data, resulting in different checkpoints. We report the best results obtained from these checkpoints.

4.2 S2TT Evaluation

The target text that is generated along with the target speech from the S2ST translation mode of SeamlessM4T, was used to compare it with the target language reference for evaluation. Instead of conventional evaluation methods like BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020)

[6] We acknowledge that the SeamlessExpressive or the large models might provide better scores. However, the SeamlessExpressive model does not support the Japanese language, and the large model requires high computational resources for fine-tuning.

[7] Instead of introducing special tokens, the labels were prepended as text.

Training Data	Fine-tuning Setting	Evaluation Data En-Ja	 En-De
-	No fine-tuning	30.28	50.47
En-Ja	w/o emotion labels	30.77	-
	w/ emotion labels	**33.18**†	-
En-De	w/o emotion labels	-	54.92
	w/ emotion labels	-	55.13
Mixed	w/o emotion labels	32.51	55.60
	w/ emotion labels	32.52	**55.84**

Table 3: BLEURT scores in percentage on the MELD-ST test sets for the S2TT experiments with different training data and fine-tuning settings. † indicates the difference in the scores of with and without emotion labels is statistically significant at $p < 0.05$.

was used to access translation quality because professional translations always differ greatly from literal interpretations, especially in languages like Japanese. Relying solely on n-gram matching for evaluation becomes challenging in such cases.

4.3 S2ST Evaluation

ASR-BLEU (Lee et al., 2022) was used to evaluate the quality of the generated target speech. We used the implementation in the Seamless Communication repository,[8] which uses Whisper (Radford et al., 2022) as the underlying model.

Additionally, the generated speeches were evaluated against the original source language speech files considering various criteria such as prosody, voice similarity, pauses, and speech rate, following Seamless Communication et al. (2023b). Details of the prosody evaluation are provided in Appendix B.

5 Results

5.1 S2TT Resuts

Table 3 shows the S2TT results. We can see that the quality of the translations generally improved after fine-tuning, and incorporating emotion labels led to slight enhancements. Using separate data or mixed data for fine-tuning does not show a significant difference.

5.2 S2ST Results

Training Data	Fine-tuning Setting	Evaluation Data En-Ja	 En-De
-	No fine-tuning	0.15	8.32
En-Ja	w/o emotion labels	0.46	-
	w/ emotion labels	0.16	-
En-De	w/o emotion labels	-	8.37
	w/ emotion labels	-	8.23
Mixed	w/o emotion labels	**0.47**	**9.82**
	w/ emotion labels	0.14	8.85

Table 4: ASR-BLEU scores on the MELD-ST test sets for the S2ST experiments with different training data and fine-tuning settings.

Table 4 shows the S2ST results. We can see that fine-tuning the SEAMLESSM4T model improves the ASR-BLEU results. However, fine-tuning with emotion labels does not help. The rest of the metrics, such as prosody similarity and vocal similarity, do not change significantly after fine-tuning.[9] The reason for this is that SEAMLESSM4T doesn't bother to learn the pronunciation features in the original speech.[10] Pauses and speed changed a bit after fine-tuning, which can be assumed to be because the translation is closer to the reference after fine-tuning.

5.3 Discussion

It is generally observed that En-De provides higher translation scores compared to En-Ja in both S2TT and S2ST. Japanese and English are very different, making the translation difficult. With the help of emotion labels, the BLEURT scores improved slightly, but not enough to be regarded as a translation with good quality. German is more similar to English and gets higher scores. After manually checking the results, most of the sentences were very clear and correct. The reason why the addition of emotion labels does not improve the results is probably because the sentences in the test set do not change much due to the difference in emotion labels.

6 Conclusion

In this study, we presented the MELD-ST dataset, an ST dataset in an emotionally rich situation, which contains both En-Ja and En-De language pairs. We conducted baseline S2TT and S2ST experiments with and without utilizing emotion labels, which showed that emotion labels can boost performance in some settings.

[8] https://github.com/facebookresearch/seamless_communication/tree/main/src/seamless_communication/cli/eval_utils

[9] The scores are presented in Appendix B Tables 5 and 6.

[10] This means that SEAMLESSM4T does not capture the pronunciation features like prosody or vocal style in the original speech as SEAMLESSEXPRESSIVE does.

For improving the translation performance, several approaches can be considered, such as training a multitask model of speech emotion recognition and ST, and utilizing dialogue context in translation.

Limitations

Some audio files in the MELD-ST dataset may contain more words than its presented text due to alignment issues. When evaluating translations of the same text with different emotion labels, it's challenging to determine the cause of differences in results. Pinpointing whether the variance stems from the emotion label or the extra information within the audio proves difficult.

The MELD-ST dataset is constructed based on acted speech, and further research is required for more natural settings such as spontaneous dialogues. As explained in Section 4.1, the models used for the experiments in this paper are basic ST models, and further performance gain could be obtained from models specifically tailored for emotion-aware speech translation.

Ethics Statements

The MELD-ST dataset will be released to the research community with restricted access to facilitate the advancement of emotion-aware speech translation, considering the risk of unintended usage of the dataset violating the copyright. In the dataset, English text, speech, and emotion labels were gathered from publicly available sources. The English data used to correct timestamps, as well as the Japanese and German text and speech, were sourced from a Blu-ray disk. Individuals seeking access to this dataset will be requested to confirm that their purpose for using it is solely for research.

Some utterances in the dataset may contain offensive contents like swear words, to the extend that is accepted by the public (i.e., to be able to appear in a TV series).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K28144 and JP23KJ1356.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C:

a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics*.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS Corpus and Massively Multilingual Speech-to-Speech Translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. In *Speech and Computer*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis.

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*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the 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.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. AudioPaLM: A Large Language Model That Can Speak and Listen.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023a. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023b. Seamless: Multilingual Expressive and Streaming Speech Translation.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shuichiro Shimizu, Chenhui Chu, Sheng Li, and Sadao Kurohashi. 2023. Towards speech dialogue translation mediating speakers of different languages. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*.

Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, and Divette Marco. 2023. Dialogs re-enacted across languages.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2023. GigaST: A 10,000-hour Pseudo Speech Translation Corpus. In *Proc. INTERSPEECH 2023*.

A Dataset Alignment Quality

We manually checked the alignment quality with the alignment process in Section 3.3 for the En-Ja part of the dataset. We sampled 307 utterances from the utterances and checked the alignment with the following criteria:

- Correct: The English and Japanese utterances match
- No translation: The English utterance is not aligned with the Japanese utterance for various reasons
- Rough segmentation: Multiple English utterances correspond to one Japanese utterance
- Others: The English utterance and the Japanese utterances do not match for other reasons

We found that 64.5% belong to "Correct," 16.6% belong to "No translation," 5.9% belong to "Rough segmentation," and 13.1% belong to "Others." For "No translation" the reasons are as follows: 1) The Japanese subtitles often omit information; 2) The translation does not appear as subtitles for simple utterances such as "Hi"; 3) In cases where one Japanese utterance corresponds to multiple English utterances, only one English utterance could be aligned to the Japanese utterance. Because these parts cannot be used in the translation experiment, such utterances were automatically detected and discarded.

B S2ST Evaluation on Prosody

Here, we provide details of the S2ST evaluation on prosody. We used the STOPES library for the evaluation.[11] Prosody similarity, measured by AUTOPCP, evaluates speech patterns, including rhythm, intonation, and stress. Vocal similarity, analyzed through cosine similarity with the function VSIM, quantifies acoustic characteristics like pitch and tone. Pauses in speech and speech rate are evaluated using local_prosody. This tool aligns audio with its corresponding text, annotates word duration, and identifies pause locations to calculate and evaluate the data.

The evaluation results on the prosody of the generated speech in the S2ST experiments are presented in Tables 5 and 6.

[11] `https://github.com/facebookresearch/stopes/tree/main/stopes/eval`

Training Data	Fine-tuning Setting	AUTOPCP	VSim	Pause	Rate
-	No fine-tuning	1.75	0.0034	0.501	-0.09
En-Ja	w/o emotion labels	1.83	-0.0004	-0.086	0.47
	w/ emotion labels	1.94	-0.0020	-0.122	0.50
Mixed	w/o emotion labels	1.88	0.0020	0.620	-0.12
	w/ emotion labels	1.89	-0.0023	0.482	0.08

Table 5: Generated Japanese target speech evaluation results.

Training Data	Fine-tuning Setting	AUTOPCP	VSim	Pause	Rate
-	No fine-tuning	2.00	0.0091	0.501	0.09
En-De	w/o emotion labels	2.07	-0.0083	0.091	0.63
	w/ emotion labels	2.05	0.0089	0.138	0.63
Mixed	w/o emotion labels	2.08	0.0082	0.477	0.07
	w/ emotion labels	2.07	0.0085	0.482	0.08

Table 6: Generated German target speech evaluation results.

C Dataset Examples

Tables 7 and 8 show some examples from the MELD-ST dataset.

D Translation Examples

Table 9 shows some observed examples from the S2TT experiments, which can show the potential of emotion labels to help improve translation quality. For the first sentence, when the model is trained without considering emotion, it translates the source language directly. However, when the emotional label is incorporated, the translated text exhibits more joy. The original text succinctly conveys the emotions without ambiguity or implication, leading to translations that remain consistent whether fine-tuned with or without the use of emotion labels.

Emotion	English	Japanese
neutral	But um, I don’t think it’s anything serious.	大したことない
surprise	Oh my God!	ヤダマジ？ウソ
surprise	Oh my God!	やったわ!
surprise	This sounds like a hernia. You have to-you—you go to the doctor!	ヘルニアだな医者へ
joy	Thank you⋯we’re so excited	ありがとう楽しみです
anger	Hey, Ross!!! I told you I don’t!	ロスいい加減にして

Table 7: Example utterances from the MELD-ST En-Ja set.

Emotion	English	German
neutral	What do you mean?	Wie meinst du das?
surprise	Are you serious?	Das kann nicht sein.
surprise	Oh my God!	Ah! Oh, mein Gott!
surprise	Oh my God!	Ich glaub’s nicht!
joy	Oh my God!	Ich glaub’s nicht!
joy	Oh crap!	So ein Käse!
joy	They taste so good.	Die sind wirklich köstlich.
anger	He does not look happy.	Er scheint nicht begeistert zu sein.
anger	I can’t believe this! This is like the worst night ever!	Das ist wirklich der schrecklichste Abend, den ich je hatte.

Table 8: Example utterances from the MELD-ST En-De set.

Input	Reference	No fine-tuning	w/o emotion labels	w/ emotion labels
This game is kind of fun.	Hey, das Spiel macht doch Spaß. (Hey, the game does make fun)	Hey, das Spiel macht doch Spaß. (The game is a bit of fun).	Das Spiel ist ein bisschen lustig. (The game is a bit of fun.)	Das Spiel ist ja wirklich lustig. (The game is really fun.)
I’ m very glad you’ re here.	Dass du da bist, macht mich sehr glücklich. (It makes me very happy that you’re here)	Ich bin sehr froh, dass du hier bist. (I’m very happy, that you are here)	Ich bin froh, dass du da bist. (I’m happy, that you are here).	Ich bin froh, dass du da bist. (I’m happy, that you are here).
You are so sweet.	うれしいわ (So happy)	あなたはとても可愛い。 (You are really cute.)	優しいわ (Gentle)	かわいい人ね (Cute person)

Table 9: Example of emotional fine-tuning in translation.