Emotion-aware Speech-to-text Translation with Generative Error Correction

Zhengdong Yang　Chenhui Chu
Kyoto University
zd-yang@nlp.ist.i.kyoto-u.ac.jp　chu@i.kyoto-u.ac.jp

Abstract

This paper explores emotion-aware speech-to-text translation (ST) using generative error correction (GER) by large language models (LLMs). Despite recent advancements in ST, the impact of the emotional content has been overlooked. First, we enhance the translation of emotional speech by adopting the GER paradigm: Finetuned an LLM to generate the translation based on the decoded *N*-best hypotheses. Next, we combine the emotion labels into the LLM finetuning process to enable the model to consider the emotion content. Experiments show that GER and the integration of emotion labels are effective on the English-Japanese language pair. This research lays the foundation for more sophisticated models that consider emotional nuances in speech.

1 Introduction

Speech-to-text translation (ST) is a task where the model takes speech in one language as input and translates it into text in another language. ST performance has greatly improved over the recent years with significant efforts on datasets [1, 2, 3, 4, 5, 6] and models [7, 8, 9, 10]. However, an essential aspect often overlooked in speech translation is the emotion of speech.

Human speech naturally includes emotions. In real-life conversations, a listener often uses cues from the speaker's voice tone to grasp what is being said. Therefore, emotion can significantly influence the results of translating speech. As the instance shown in Figure 1, the phrase "I can't believe this" can convey a range of emotions, from surprise and shock to awe and excitement, which can alter its translation in another language. In Japanese, the translation might vary from "信じられない" (expressing surprise) to "どうしてこんなことに" (expressing frustration).

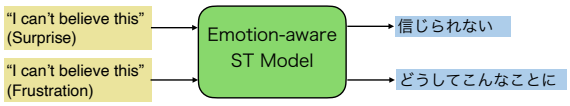

Figure 1 The expectation for an emotion-aware ST model. It can generate appropriate translation based on the emotion of the input speech.

Emotion has been studied in machine translation (or text-to-text translation) studies [11] and other tasks in natural language processing (NLP), such as sentiment analysis and recognizing emotions in conversations [12]. However, there has been little effort focusing on emotion in ST. Seamless Expressive [8] examines the preservation of emotional states in speech-to-speech translation, without addressing the influence of emotions on the semantic aspects of translation. Chen et al. [13] constructed the MELD-ST dataset for emotion-aware ST, but further community effort investigating the methodology for this task is required.

Meanwhile, recent advancements in large language models (LLMs) leads to growing interest in leveraging their capabilities in modalities beyond text including speech. Training end-to-end ST models often face challenges due to insufficient speech-text parallel data. However, LLMs are trained on vast amounts of textual data and obtain powerful textual generation abilities, which can enhance the ST performance. This has been proven by recent studies that use LLMs as decoders for ST systems [14] or as Generative Error Correction (GER) models to improve ST qualities [15].

Speech-text parallel data is scarce, and it is even scarcer when it includes emotion annotations. Therefore, leveraging external models like LLMs to help the system understand the correlation between emotion and language can be greatly beneficial. However, to the best of our knowledge, there have not been studies on utilizing LLMs for emotion-aware ST. Therefore, this research aims to pio-

neer the exploration of the effectiveness of emotion-aware ST by: (a) adopting the LLM GER paradigm, (b) adding emotion labels into the GER finetuning process. We will introduce these two proposals in detail in the next section.

2 Method

2.1 Generation Error Correction

As shown in Figure 2, the GER paradigm consists of two main parts: a pre-trained ST model for generating N-best hypotheses, and an LLM finetuned to work as a GER model to re-generate the translation prediction.

2.1.1 N-best Hypotheses Generation

In order to generate inputs for the LLM GER model, A pre-trained ST model is utilized to decode N-best hypotheses from input speech with beam search. More specifically, given an input speech S in source language, the ST model translates it into target language text by beam search decoding with a beam size of M, which generates N-best hypotheses list $\mathcal{T}_N = \{T_1, T_2, ..., T_N\} (N \leq M)$. In practice, we set $N = M$. The list serves as the preliminary prediction and a part of the input for the LLM GER model.

2.1.2 GER Finetuning

Inspired by the methods proposed by [15], we leverage LLMs to generate a final translation result based on the decoded N-best hypotheses. We expect our model can utilize the strong linguistic and reasoning ability of LLMs to integrate the rich information in the inputs to generate a higher-quality translation result. This new generative paradigm can be formulated as:

$$T = M_{EST}(\mathcal{T}_N, I) \tag{1}$$

where I is a proper instruction for LLM prompting. The goal of our model is to learn a mapping M_{EST} from N-best hypotheses to the true translation. Following the typical sequence-to-sequence learning strategy, we employ the ground-truth translation T^* as the supervision signal and optimize the LLM to learn M_{EST} in an auto-regressive manner. The cross-entropy-based training loss is defined as:

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(t_l^* | t_{l-1}^*, ..., t_1^*; \mathcal{T}_N, I) \tag{2}$$

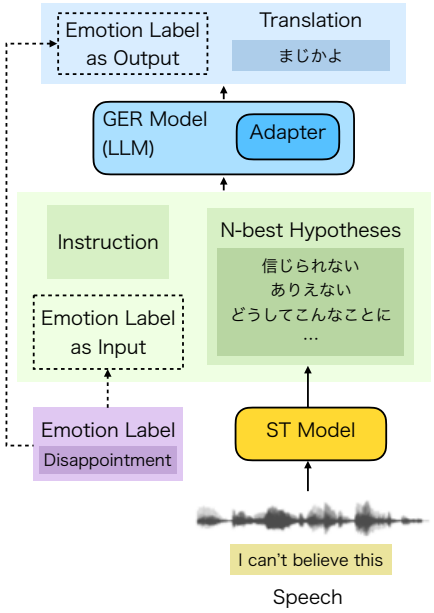

Figure 2 Overview architecture of our proposed model.

where t_l^* is the l-th token of T^*, L denotes the sequence length, and θ denotes the learnable parameters in LLM (i.e., adapter).

2.1.3 Parameter-Efficient Finetuning

Considering the large model size of LLMs, we adopt the efficient finetuning strategy LLaMA Adapter [16]. It inserts a set of learnable adaptation prompts into the top-L of total H Transformer layers [17] in a pretrained LLM to learn high-level semantics. We denote the prompt for l-th layer as $P_l \in \mathbb{R}^{U \times D}$, where U is the prompt length and D is embedding size.

Assume we have M tokens $T_l \in \mathbb{R}^{M \times D}$ including instruction and already generated response, now we aim to predict the $(M+1)$-th token as the response. The learnable adaptation prompt is concatenated with T_l as the prefix, i.e., $[P_l; T_l] \in \mathbb{R}^{(U+M) \times D}$, which provides learned instruction knowledge to guide the subsequent response generation.

Furthermore, considering the prompt P_l is randomly initialized and thus could disturb the LLM tuning at the early training stage, a zero-initialized attention mechanism is devised to mitigate such disturbance.

2.2 Integration of the Emotion Labels

We incorporate emotion labels into the GER fine-tuning process to investigate how emotional content influences translation outcomes. We evaluate the following two ap-

Table 1 Statistics of MELD-ST dataset for different language pairs (Lang.) and splits. There are 7 types of emotion labels: Neutral (Neu.), Joy (Joy.), Sadness (Sad.), Fear (Fea.), Anger (Ang.), Surprise (Sur.), Disgust (Dis.); and 3 types of sentiment labels: Neutral (Neu.), Positive (Pos.), Negative (Neg.)

Lang.	Split	Total	Neu.	Joy.	Sad.	Fea.	Ang.	Sur.	Dis.	Neu.	Pos.	Neg.
en-ja	Train	8,069	3,836	1,284	603	209	982	917	238	2,518	1,715	3,836
	Validation	1,008	482	176	84	31	116	97	22	482	229	297
	Test	1,008	479	186	73	25	85	121	39	479	253	276
en-de	Train	9,314	4,402	1,571	656	232	1,096	1,096	261	4,402	2,084	2,828
	Validation	1,164	550	202	99	31	127	130	25	550	271	343
	Test	1,164	550	218	92	32	102	131	39	550	288	326

proaches:

2.2.1 Emotion Labels as GER Inputs

We configure the GER model to generate translations based not only on the decoded N-best hypotheses but also on the ground-truth emotion labels. This approach allows us to assess the upper bound of the improvement by incorporating emotional content. Then the paradigm can be formulated as:

$$T = M_{EST}(E, \mathcal{T}_N, I) \tag{3}$$

where E is the emotion label. The cross-entropy-based training loss is defined as:

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(t_l^* | t_{l-1}^*, ..., t_1^*; E, \mathcal{T}_N, I) \tag{4}$$

2.2.2 Emotion Labels as GER Outputs

In practice, ground-truth emotion labels are unavailable, necessitating their prediction. We propose using the GER model to directly predict these labels. Consequently, based on the hypotheses, the model first generates emotion labels and then the translation. This approach can be considered multitask learning for the GER model. Then paradigm will be:

$$O_{E,T} = M_{EST}(\mathcal{T}_N, I) \tag{5}$$

where $O_{E,T}$ is the concatenated sequence of E and T. The cross-entropy-based training loss is:

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(o_l^* | o_{l-1}^*, ..., o_1^*; \mathcal{T}_N, I) \tag{6}$$

where o_l^* is the l-th token of the ground truth of $O_{E,T}$.

3 Experiments

3.1 Dataset

In this study, we use the MELD-ST dataset [13], an ST dataset in an emotionally rich situation, which contains both English-Japanese and English-German language pairs. The dataset is constructed from translations obtained from a Blu-ray disk of TV series *Friends* and emotion labels from the MELD dataset [18].

As in MELD, the utterances are labeled with 7 different emotions and 3 different sentiments. We added both types of labels into the LLM instructions in our experiments. The dataset statistics are summarized in Table 1.

3.2 Settings

Models used for different parts in our proposed architecture includes:

- **ST Model**: We use the state-of-the-art SeamlessM4T2 [7], a Transformer-based model that supports speech-to-text translation for up to 100 languages. Experiments are conducted with two model sizes: medium and large.
- **LLM GER Model**: We select the popular LLaMA-2 [19] for our architecture.
- **Adapter**: We follow the default settings of LLaMA Adapter [16]. The number of tunable Transformer layers L is set to $H - 1$, which means all layers except the first one are tunable with inserted prompts. The prompt length U is set to 10.

The batch size is set to 4, with accumulation iterations set to 8 (i.e., the real batch size is 32). We train for 2 epochs with the AdamW optimizer [20], with the learning rate initialized at $1e^{-2}$ and then linearly decreased to $1e^{-5}$

Table 2 ST results on MELD-ST dataset.

Language Pair	ST Model Size	GER Model	Emotion Labels	BLEU	BERTScore	BLEURT
en-ja	Medium	no	no	2.14	71.7	**28.9**
		yes	no	2.50	74.0	25.6
		yes	GER Input	**3.50**	**74.1**	26.5
		yes	GER Output	2.90	73.9	25.2
	Large	no	no	1.95	71.0	**28.3**
		yes	no	3.02	**74.1**	26.4
		yes	GER Input	**3.58**	**74.1**	25.8
		yes	GER Output	3.43	73.7	25.3
en-de	Medium	no	no	10.25	75.8	50.3
		yes	no	10.02	76.7	52.2
		yes	GER Input	10.13	76.7	52.2
		yes	GER Output	**10.54**	**76.9**	**52.7**
	Large	no	no	**11.73**	76.2	52.7
		yes	no	10.96	77.0	54.0
		yes	GER Input	11.28	**77.1**	**54.3**
		yes	GER Output	11.07	76.8	53.5

during training.

We conducted experiments for two approaches for adding the emotion labels: adding them as GER inputs to provide an ideal scenario that represents the performance upper bound, and adding them as GER outputs to provide a more practical scenario. The proposed approaches are compared with two baselines: one using the LLM GER model without emotion labels and one without using the LLM GER model at all.

3.3 Results

The results are presented in Table 2. To evaluate the quality of translations, we use several evaluation metrics, including BLEU, BERTScore, and BLEURT.

For the en-ja language pair, BLEU scores indicate that the GER model shows an improvement over the original SeamlessM4T model when both ST model sizes are considered. Incorporating emotion labels further enhances this improvement. Using emotion labels as inputs to the GER model provides a performance upper bound while predicting these labels with the GER model results in a slight reduction in performance gains. Although this pattern aligns with our initial assumptions, it does not hold across other evaluation metrics, such as BERTScore and BLEURT.

Conversely, for the en-de language pair, we fail to observe similar trends, suggesting that the proposed method does not apply universally across language pairs. A potential explanation for this discrepancy is the greater cultural differences between English and Japanese compared to English and German, which might make emotions more influential in translations between the former pair.

Upon reviewing the generated N-best list, we infer that the lack of diversity in the list (e.g. a list of “ほんとに”, “本当に ??”, “本当に”, “ホントに”, and “本当に ?”). may limit the GER model’s ability to select appropriate translations, even when taking emotional information into account. This leads to the future direction of increasing the diversity of the N-best list (e.g. by using temperature-based sampling).

4 Conclusion

In this paper, we pioneer the investigation of emotion-aware ST using LLMs. We propose to adopt the GER method and integrate emotion labels into the GER finetuning process. The experimental results show the effectiveness on certain language pairs. We propose several potential future directions to improve our method including increasing the diversity of the N-best list and injecting more acoustic information from the speech into the GER finetuning process.

Acknowledgment

This work was supported by the SPRING program of Kyoto University and JSPS KAKENHI Grant Number JP23K28144.

References

[1] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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)**, 2019.

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

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

[4] 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. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In **Proc. Interspeech 2021**, 2021.

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

[6] Agarwal et al. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In **Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)**, 2023.

[7] Seamless Communication et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation, 2023.

[8] Seamless Communication et al. Seamless: Multilingual Expressive and Streaming Speech Translation, 2023.

[9] Paul K. Rubenstein et al. AudioPaLM: A Large Language Model That Can Speak and Listen, 2023.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

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

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

[13] Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. Meld-st: An emotion-aware speech translation dataset. **arXiv preprint arXiv:2405.13233**, 2024.

[14] Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. On decoder-only architecture for speech-to-text and large language model integration. In **2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, pp. 1–8. IEEE, 2023.

[15] Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. Gentranslate: Large language models are generative multilingual speech and machine translators. **arXiv preprint arXiv:2402.06894**, 2024.

[16] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. **arXiv preprint arXiv:2303.16199**, 2023.

[17] A Vaswani. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017.

[18] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. **arXiv preprint arXiv:1810.02508**, 2018.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.