INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Audio-Visual Praise Estimation for Conversational Video based on Synchronization-Guided Multimodal Transformer

Nobukatsu Hojo, Saki Mizuno, Satoshi Kobashikawa, Ryo Masumura, Mana Ihori, Hiroshi Sato, Tomohiro Tanaka

NTT Computer and Data Science Laboratories, NTT Corporation, Japan
nobukatsu.hojo@ntt.com

Abstract

This study investigates praise estimation, the task of estimating the existence of preferable behaviors of a speaker in a conversational video. To estimate praises from multimodal information, considering synchronized behavior across modalities is important. Such cross-modal synchronization can be modeled by the conventional multimodal Transformer in a time-axis concatenation architecture because it models relevance between all time steps of all input modalities using attention matrices. However, the attention matrices are so high-dimensional that the model training can be difficult with a limited amount of training data. To alleviate this problem, we propose introducing a loss function representing the prior knowledge that the attention should link around the synchronized time steps across the input modalities. Our experiments on a business negotiation conversation corpus showed that the proposed method could improve the praise estimation's macro F1.

Index Terms: conversation support system, multimodal analysis, multimodal Transformer, synchronization

1. Introduction

Due to the impact of COVID-19, daily conversations have been shifting from face-to-face to online. Since it is relatively easy to store data on online conversations and for machines to intervene in such interactions, conversation support systems [1–6] can be more prevalent. They often visualize and give feedback on verbal and non-verbal information such as fillers, facial emotions, and speech duration. Estimating the existence of preferable behaviors of a speaker ("praise estimation" in this study) and giving feedback on them would also encourage self-reflection and further improvement of conversational skills. While previous studies have proposed suggestion feedback systems [5, 7, 8], they were rule-based and required high costs to build and extend. In this regard, this study tackles machine learning-based praise estimation for utterances in conversational videos. In particular, we aim to estimate praises for a seller's utterance in a dyadic business negotiation conversation between a seller and a buyer.

Praising is regarded as identifying the high conversational skills observed in an utterance. The conversational skills are related to multimodal information of the speaker [9–11]. The listener's behavior, such as facial expressions, may also suggest the quality of the speaker's utterance. To consider such information, the model in this study estimates praises from three modalities, the speaker's speech, the speaker's video, and the listener's video, for the duration of a target speaker's utterance (Fig. 1). Previous studies have also shown that considering synchronized behavior across modalities (cross-modal synchronization), such as speaking with/without gestures or smiling simultaneously, is important to estimate skills from conversation data [10, 12]. This suggests that an important relationship for praise estimation lies in synchronized time steps across the three modalities.

Previous studies have shown that Transformer can effectively model multimodal information in various tasks, including personality and skill estimation from conversation data [13–15]. We investigate a multimodal Transformer for praise estimation, in particular, based on a time-axis concatenation architecture [16]. This architecture concatenates multimodal features along the time axis to form an input sequence to a Transformer layer. The advantage is that an attention matrix in a Transformer has the potential to flexibly model the inherent relevance between all time steps of all modalities. However, model training can be difficult with a limited amount of training data. This is because numerous relationships, i.e., combinations between all time steps of all modalities, need to be modeled solely from training data.

The key idea of the proposed method is to utilize prior knowledge as well as data for training a multimodal Transformer. In particular, the proposed method constrains the attention matrix to reflect the prior knowledge that there is a meaningful relationship around synchronized time steps across modalities. This idea is inspired by previous studies in speech processing [17, 18] and language processing [19]. For voice conversion [17] and text-to-speech [18], they introduced a loss function that guides an attention matrix to be diagonal. This loss function reflects the prior knowledge that temporal alignment between inputs and outputs should be monotonic and nearly linear. For language modeling, Longformer [19] also constrains the attention matrix so that several specific tokens have global attention while others have local windowed ones. This constraint reflects prior knowledge about the tasks and input tokens. For a multimodal Transformer, the proposed method introduces a loss function to guide the attention matrix to link around the synchronized time steps across modalities. Our experiments on praise estimation demonstrate that the proposed method improves the performance of multimodal Transformers.

2. Related Works

In addition to time-axis concatenation, there are several major architectures of a Transformer that can incorporate multimodal inputs; late fusion [20], feature-axis concatenation [15, 21], early summation [22,23], and cross-attention [24–26]. The late-fusion model integrates multimodal information at the decision level. It encodes each modality with a different Transformer and cannot incorporate temporal relationships across modalities. The feature-axis concatenation model concatenates multimodal features along the feature axis. Then, the concatenated vectors are input to a Transformer layer. Similarly, the early summation model adds multimodal features to construct an input vector to a Transformer. These two models are less flexible than the time-axis concatenation because all modalities share common attention matrices. Cross-attention uses one modality to estimate the relevance of each position in another modality. It enables flexible modeling similar to time-axis concatenation. For cross-attention, our approach to guide an attention matrix to link around synchronized time steps across modalities has yet to be investigated. While the proposed method is formulated on the basis of the time-axis concatenation, our approach can be similarly applied to cross-attention models.

10.21437/Interspeech.2023-1464

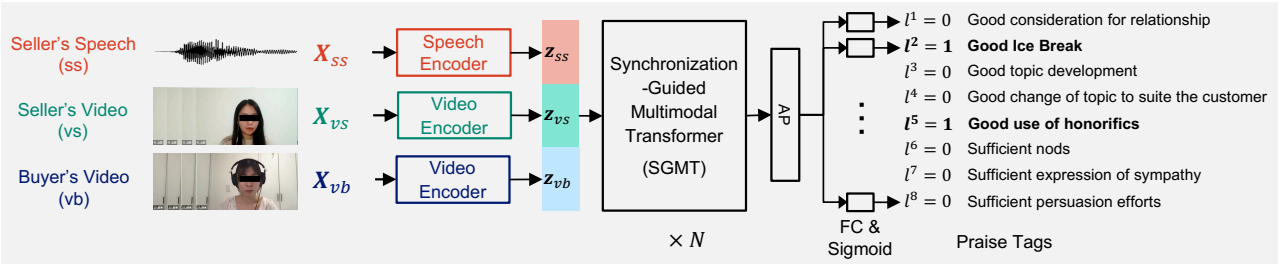


Figure 1: *Schematic diagram of the praise estimation task and proposed method. "AP" and "FC" denote an attention pooling and fully-connected layer, respectively.*

3. Praise Estimation Method

3.1. Praise Estimation Task

Let $\boldsymbol{X} = \{\boldsymbol{X}_{\rm ss}, \boldsymbol{X}_{\rm vs}, \boldsymbol{X}_{\rm vb}\}$ be a data stream of a seller's speech (ss), seller's video (vs), and buyer's video (vb) that correspond to the time interval of a target seller's utterance, respectively. Let $\boldsymbol{l} = \{l^1, \cdots, l^K\}$ be the praise tags corresponding to the target seller's utterance, where K is the number of praise tags and $l^k \in \{0, 1\}$ denotes that the utterance is {"not praised", "praised"} regarding the k-th praise tag, respectively. Figure 1 illustrates the eight praise tags used in this study. Since multiple praise tags can correspond to an utterance, the praise estimation is defined as the multi-label classification task, which determines $\boldsymbol{l}$ from

$$\hat{\boldsymbol{l}} = f(\boldsymbol{X}_{\rm ss}, \boldsymbol{X}_{\rm vs}, \boldsymbol{X}_{\rm vb}; \Theta), \tag{1}$$

where $\hat{\boldsymbol{l}}$ is the predicted praise tags, $f(\cdot)$ is the classification function determined by the model, and Θ is a parameter set of the model.

3.2. Baseline Method

The baseline method is based on the time-axis concatenation architecture. It first uses pre-trained encoders to extract features of each modality $\boldsymbol{Z}_{\rm ss}, \boldsymbol{Z}_{\rm vs}, \boldsymbol{Z}_{\rm vb}$ from $\boldsymbol{X}$,

$$\boldsymbol{Z}_{\rm ss} = \texttt{SpeechEncoder}(\boldsymbol{X}_{\rm ss}; \theta_s) \tag{2}$$
$$\boldsymbol{Z}_{\rm vs} = \texttt{VideoEncoder}(\boldsymbol{X}_{\rm vs}; \theta_v) \tag{3}$$
$$\boldsymbol{Z}_{\rm vb} = \texttt{VideoEncoder}(\boldsymbol{X}_{\rm vb}; \theta_v), \tag{4}$$

where $\texttt{SpeechEncoder}(\cdot)$ and $\texttt{VideoEncoder}(\cdot)$ are a projection function from data to the feature vector for speech and video, respectively. θ_s and θ_v are parameters of the encoders. $\boldsymbol{Z}_m \in \mathbb{R}^{D \times T_m}$ is the feature vector of modal $m \in \{\rm ss, vs, vb\}$ where D and T_m are the feature dimension and time length of the feature of modality m, respectively. Note that the estimation models in this study do not use textual information, such as speech recognition results. However, several praise labels, such as honorifics and ice-breakers, are related to textual information. This model design is based on the study that showed pre-trained speech encoders can implicitly capture textual information from audio data [27].

The feature vectors are then concatenated along the time axis and fed to the Transformer layers. The n-th Transformer encoder block comprises the n-th hidden representations $\boldsymbol{S}^{(n)}$ from the lower inputs $\boldsymbol{S}^{(n-1)}$ as

$$\boldsymbol{S}^{(0)} = [\boldsymbol{Z}_{\rm ss}; \boldsymbol{Z}_{\rm vs}; \boldsymbol{Z}_{\rm vb}] \tag{5}$$
$$\boldsymbol{S}^{(n)} = \texttt{Transformer}(\boldsymbol{S}^{(n-1)}; \theta_{\rm T}^{(n)}), \tag{6}$$

where $\texttt{Transformer}(\cdot)$ is the Transformer encoder block [28] in the pre-layer normalization architecture [29]. $\theta_{\rm T}^{(n)}$ is the model parameter of the n-th Transformer layer. The hidden representation is projected to the posterior probability of the praise tags,

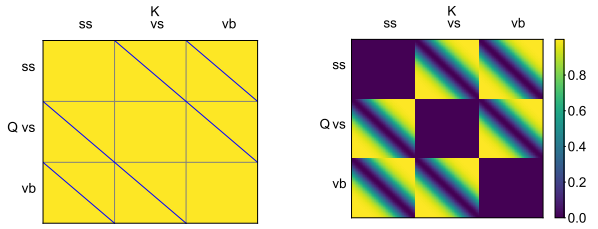
$$P(l^k|\boldsymbol{X}) = \texttt{sigmoid}(\texttt{FC}(\texttt{AP}(\boldsymbol{S}^{(N)}; \theta_{\rm AP}); \theta_{\rm FC}^{(k)})), \tag{7}$$

where $\texttt{sigmoid}(\cdot), \texttt{FC}(\cdot), \texttt{AP}(\cdot)$ are a sigmoid, a fully-connected, and an attention pooling layer, respectively. $\theta_{\rm AP}$ and $\theta_{\rm FC}^{(k)}$ are model parameters of the attention pooling layer and a fully-connected layer corresponding to the k-th label, respectively.

The model parameters $\Theta = \{\theta_s, \theta_v, \{\theta_{\rm T}^{(n)}\}_n, \theta_{\rm AP}, \{\theta_{\rm FC}^{(k)}\}_k\}$ are optimized by minimizing cross-entropy loss $\mathcal{L}_{label}$ using training data $\mathcal{D}$,

$$\mathcal{L}_{label} = \sum_{\{\boldsymbol{X}, \boldsymbol{l}\} \in \mathcal{D}} \sum_k -\log P(l^k|\boldsymbol{X}; \Theta). \tag{8}$$

Note that speech and video encoders are pre-trained and frozen during training.



(a) *Synchronized time steps across modalities (blue lines) in an attention matrix.*

(b) *Example of attention penalty matrix $\boldsymbol{W}$ used for calculating synchronization-guided loss.*

Figure 2: *Illustration of proposed method. For simplicity, we show the examples when $T_{\rm ss} = T_{\rm vs} = T_{\rm vb}$.*

3.3. Proposed Synchronization-Guided Multimodal Transformer (SGMT)

A Transformer layer mainly consists of a multi-head attention layer and a position-wise feed-forward network. In each of the multi-heads of each Transformer layer, the attention matrix $\boldsymbol{A}$ represents the relevance within the input sequence. Therefore, we introduce a constraint on $\boldsymbol{A}$ based on the prior knowledge that there is a meaningful relationship around synchronized time steps across modalities.

The attention matrix $\boldsymbol{A}$ is defined as a scaled dot product of a query matrix $\boldsymbol{Q}$ and a key matrix $\boldsymbol{K}$. $\boldsymbol{Q}$ and $\boldsymbol{K}$ are calculated by multiplying weight matrices to a token of each time step of the input sequence. Therefore, in the time-axis concatenation architecture, the query and key matrices consist of small

Figure 3: *Photo of participants engaged in negotiation task*

matrices, each of which corresponds to a modality as follows,

$$\boldsymbol{Q} = [\boldsymbol{Q}_{\rm ss}; \boldsymbol{Q}_{\rm vs}; \boldsymbol{Q}_{\rm vb}] \tag{9}$$

$$\boldsymbol{K} = [\boldsymbol{K}_{\rm ss}; \boldsymbol{K}_{\rm vs}; \boldsymbol{K}_{\rm vb}], \tag{10}$$

where $\boldsymbol{Q}_m, \boldsymbol{K}_m \in \mathbb{R}^{d \times T_m}$ for each modality m and d denotes the vector dimension of the queries and keys. The scaled dot product of them comprises the following attention matrix consisting of block matrices,

$$\boldsymbol{A} = \texttt{softmax}\left(\frac{\boldsymbol{Q}^\top \boldsymbol{K}}{\sqrt{d}}\right) \tag{11}$$

$$= \begin{pmatrix} \boldsymbol{A}_{\rm ss,ss} & \boldsymbol{A}_{\rm ss,vs} & \boldsymbol{A}_{\rm ss,vb} \\ \boldsymbol{A}_{\rm vs,ss} & \boldsymbol{A}_{\rm vs,vs} & \boldsymbol{A}_{\rm vs,vb} \\ \boldsymbol{A}_{\rm vb,ss} & \boldsymbol{A}_{\rm vb,vs} & \boldsymbol{A}_{\rm vb,vb} \end{pmatrix}, \tag{12}$$

where $\texttt{softmax}(\cdot)$ is a softmax function. A block matrix $\boldsymbol{A}_{m_1,m_2} \in \mathbb{R}^{T_{m_1} \times T_{m_2}}$ can be regarded as an attention from a modality m_1 to another m_2.

Cross-modal attentions are represented by non-diagonal block matrices, $\boldsymbol{A}_{m_1,m_2} (m_1 \neq m_2)$. Their diagonal components correspond to synchronized time steps across modalities (see Fig. 2a). Therefore, for each of $\boldsymbol{A}_{m_1,m_2}$, the diagonal region should be dominant. We introduce the following synchronization-guided loss to penalize $\boldsymbol{A}_{m_1,m_2}$ for not having a diagonally dominant structure. First, we define the following attention penalty matrix $\boldsymbol{W}$,

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{W}^{\rm ss,vs} & \boldsymbol{W}^{\rm ss,vb} \\ \boldsymbol{W}^{\rm vs,ss} & \boldsymbol{0} & \boldsymbol{W}^{\rm vs,vb} \\ \boldsymbol{W}^{\rm vb,ss} & \boldsymbol{W}^{\rm vb,vs} & \boldsymbol{0} \end{pmatrix} \tag{13}$$

$$W_{i,j}^{m_1,m_2} = 1 - \exp\left(-\frac{(i/T_{m_1} - j/T_{m_2})^2}{2\sigma^2}\right), \tag{14}$$

where $W_{i,j}^{m_1,m_2}$ is (i,j)-th element of block matrix $\boldsymbol{W}^{m_1,m_2} \in \mathbb{R}^{T_{m_1} \times T_{m_2}}$, which was formulated referring to the "diagonal attention loss" in previous works [17, 18]. σ is a hyperparameter. The matrix $\boldsymbol{W}$ has positive values in non-diagonal regions in the non-diagonal (cross-modal) blocks (see Fig. 2b). The synchronization-guided loss $\mathcal{L}_{sg}$ is calculated as the sum of the losses for each of multi-heads $h \in \{1, \cdots, H\}$ in each Transformer layers $n \in \{1, \cdots, N\}$,

$$\mathcal{L}_{sg} = \frac{1}{HN} \sum_{h,n} \frac{1}{T^2} ||\boldsymbol{W} \odot \boldsymbol{A}_{h,n}||_1, \tag{15}$$

where $T = T_{\rm ss} + T_{\rm vs} + T_{\rm vb}$ is the total length of the input sequence. The proposed model is trained to minimize the following loss function,

$$\mathcal{L} = \mathcal{L}_{label} + \lambda_{sg} \mathcal{L}_{sg}, \tag{16}$$

where $\lambda_{sg} \geq 0$ is a regularization parameter, which weighs the importance of $\mathcal{L}_{sg}$ relative to $\mathcal{L}_{label}$.

4. Dataset

For the experiments in this study, we used the online audio-visual negotiation corpus constructed in a previous study [30].

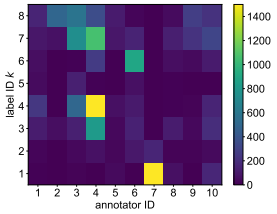


Figure 4: *Frequency of annotation of each praise label by annotators.*

Table 1: *Frequency of each label k in the dataset (8469 utterances in total).*

k	$\#l^k = 1$	ratio (%)
1	2216	26.2
2	320	3.8
3	1170	13.8
4	1898	22.4
5	194	2.3
6	839	9.9
7	1892	22.3
8	1200	14.2

While the original corpus contained 48 dyadic business negotiation conversations, we conducted additional recordings and used 64 dialogues. The conversations concerned four products (chat tools, insurance, TVs, and cell phones), and they negotiated prices, delivery dates, quantities, services, etc. There were two experienced salespeople for each of the four products, eight in total. Four men and four women were sellers, and eight men and eight women were buyers. Their ages were between 19 and 60. They gave their written informed consent before starting the experiment. The datasets are not publicly available because participants recruited in this study did not give their consent to their raw data being publicly shared.

All participants were Japanese, and the recorded conversations were in Japanese. The corpus is composed of 1027 minutes of recordings (average duration: 16.04 minutes per conversation). Video and audio were recorded during the business negotiation using a camera and microphone installed on the client's PC. The video was recorded at 25 fps with a 1280 $\times$ 720 resolution. Camera views were frontal and recorded the upper part of the body (see Fig. 3). The audio was recorded at 32 kHz. The recorded video and audio were synchronized. Participants were instructed to use headphones.

The seller's speech was segmented into inter-pausal units (IPUs) [31] using a voice activity detection technique. To ensure sales expertise is reflected in the annotations, we hired ten annotators with at least three years of sales supervisory experience. They watched the recorded conversational video and assigned eight praise tags (Fig. 1) to each utterance by the seller considering conversational contexts.

Figure 4 shows each annotator's frequency of annotation of each praise label. We can see a significant bias in the frequency of label assignments depending on the annotator. On the basis of our informal data observation, we also judged that each annotator assigned praise labels with some consistency of criteria. These suggest that although the praise labels were assigned using different criteria depending on the annotator, each assignment had validity to a certain degree. In this regard, we defined the ground-truth label as "praised" (i.e., $l^k = 1$) for an utterance when any of the ten annotators judged as such. We also defined the ground-truth label as "not praised" (i.e., $l^k = 0$) when all annotators judged as such. Table 1 shows the frequency of each label k in the dataset.

5. Experiments

5.1. Setup

To evaluate the effectiveness of the proposed method, we compared the performance of the proposed method with four conventional Transformer architectures.

- SPEECH: Unimodal Transformer using only seller's speech modality.
- LATE-FUSION: Transformer in the late-fusion architecture that combines outputs from three unimodal Transformers.
- FEATURE-CONCAT: Transformer that concatenates multimodal input features along the feature axis.
- MT (Baseline): Multimodal Transformer in the time-axis

concatenation architecture in Sec. 3.2.

- SGMT (Proposed): The proposed synchronization-guided multimodal Transformer.

SPEECH used only the seller's speech data for estimation. We set $\boldsymbol{S}^{(0)} = \boldsymbol{Z}_{\rm ss}$ in (5) in the baseline method. LATE-FUSION models each modality with a different Transformer. In (7), the output vectors from the $\texttt{AP}(\cdot)$ layer for the three modalities are added and input to the $\texttt{FC}(\cdot)$ layer. FEATURE-CONCAT concatenates feature vectors of three modalities along the feature axis. We set $\boldsymbol{S}^{(0)} = [\boldsymbol{Z}_{\rm ss}^\top; \boldsymbol{Z}_{\rm vs}^\top; \boldsymbol{Z}_{\rm vb}^\top]^\top$ in (5). Since the frame shift of video was longer than that of speech, we equalized the time length of features among modalities before the concatenation by repeating each frame of $\boldsymbol{Z}_{\rm vs}$ and $\boldsymbol{Z}_{\rm vb}$.

Pre-processing: For the acoustic features, we extracted 80 log Mel-scale filterbank coefficients. The frameshift was 10 ms. For the visual features, we detected face regions in each input frame with YOLOv3 [32] trained on the Wider Face dataset [33]. The face images were cropped, resized to 128×128, and downsampled to 2.5 fps.

Encoder configurations: For the speech encoder, acoustic features were passed through two convolution and max pooling layers with a stride of 2, so we downsampled them to 1/4 along with the time axis. We stacked six Transformer encoder blocks. All components in the speech encoder were pre-trained with end-to-end automatic speech recognition tasks using over 10K hours of speech. For the video encoder, the CNN function was composed on the basis of MobileNetV3 [34]. After that, we stacked two Transformer encoder blocks. For each Transformer block, we set the dimensions of the outputs to 128, the dimensions of the inner outputs in the position-wise feed-forward networks to 512, and the number of heads in the multi-head attention to 4. We used the Swish activation for the position-wise feed-forward networks. We pre-trained the CNN component in the video encoder through two steps. It is pre-trained with a face recognition task using VGGFace2 [35] in the first step, and a still-image-based facial expression recognition task using FER [36], RAF-DB [37], and AffectNet [38] datasets in the second step.

Multimodal Transformer configurations: For multimodal Transformers, we set the dimensions of the outputs and the inner outputs in the position-wise feed-forward network to 512. We set the number of heads in the multi-head attention to 4. We used the Swish activation function. We used a single Transformer layer (i.e., $N = 1$). We set the mini-batch size to 4 and the dropout rate in the Transformer blocks to 0.1. We used Adam [39] for optimization. We stopped the training steps based on early stopping utilizing the validation set. For the proposed model, we set the hyperparameters $\sigma = 0.2$ and $\lambda_{SG} = 4500$. These values were experimentally selected from $\sigma \in \{0.1, 0.2, 0.3, 0.7\}$ and $\lambda_{SG} \in \{45, 450, 4500\}$.

Evaluation: We excluded utterances when the face region detection failed for either speaker. This resulted in 8469 utterances for training and testing in the experiments. We used an 8-fold cross-validation method for training and testing so that the test set does not include any participants in the training set. We evaluated the overall performance by macro F1. We also confirmed the trend in Precision and Recall averaged over eight labels. To eliminate the effect of randomness in model training, we repeated the evaluation five times to calculate an average score for each experimental condition.

5.2. Results

Table 2 shows the evaluation results. We observe that the macro F1 of the proposed method was higher than that in the other conventional methods by 0.021. This demonstrated the effectiveness of the proposed method. When we compare the macro F1 of the four conventional methods, we observe that LATE-FUSION was better than SPEECH. We also observed that FEATURE-CONCAT and MT were even worse than SPEECH. These demonstrate that the baseline methods that concatenate multimodal features had difficulty estimating when utilizing multimodal information. The proposed method made the model training efficient, which led to the effective utilization of multimodal information.

Table 2: *Evaluation results of the praise estimation experiments averaged over the eight praise labels.*

	Precision	Recall	macro F1
SPEECH	.430	.170	.237
LATE-FUSION	.450	.181	.245
FEATURE-CONCAT	.391	.123	.174
MT (Baseline)	**.502**	.157	.228
SGMT (Proposed)	.465	**.190**	**.266**

We also see that Precision was higher than Recall in all conditions. Since the "praised" tags were fewer than the "not praised" tags as shown in Tab. 1, obtaining a high Recall in the experiments was difficult. We also observe that improvement by the proposed method was clear in Recall rather than Precision. The proposed method can consider multimodal information, including cross-modal synchronization in an utterance. Therefore, it can assign praise tags to a wider variety of utterances than the baseline method. This is why the proposed method could improve Recall and macro F1.

6. Discussion

We investigated praise estimation, the task of estimating the existence of preferable behaviors of a speaker in a conversational video. We proposed introducing a loss function representing the prior knowledge that the attention should link around the synchronized time steps across the input modalities. Our experiments on a business negotiation conversation corpus showed that the proposed method could improve the praise estimation's macro F1 by 0.021.

We investigated utterance-level praise estimation rather than session-level. This design was based on the expectation that users would be more self-reflective and motivated if they could recognize the specific target utterance of each praise by the system. The agreement value of the utterance-level praise tags between annotators was low, which can degrade the estimation accuracy and reproducibility of the experiments. This was in contrast to a previous study that showed a sufficient agreement value when they annotated communication skills at sub-session level (5 to 7 minutes) [40]. This is because differences in judgment criteria among annotators significantly impact the results, especially when the annotation target is short and its information is scarce. In other words, there is a trade-off between the time resolution of estimation and validity of the ground truth. Future work includes an optimal design of praise estimation considering the trade-off.

There was a limit to Recall in the experiments. A possible reason for this was the label imbalance of the praise tags. Considering label imbalance by objective function [41] can improve Recall. Another possible reason was the need for more information considered by the model. The annotators may assign praise tags considering the conversational context, including the buyer's speech, as well as the seller's speech, and the seller's and buyer's videos. Another future work considers conversational context [42] by the proposed model.

We evaluated the proposed method for the praise estimation task using a private dataset. The proposed method can also be effective in other tasks where cross-modal synchronization is important, such as multimodal emotion recognition and personality estimation from conversational videos. Future work includes evaluating the proposed method for such tasks using publicly available datasets.

7. References

[1] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard *et al.*, "The tardis framework: intelligent virtual agents for social coaching in job interviews," in *International Conference on Advances in Computer Entertainment Technology*. Springer, 2013, pp. 476–491.

[2] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 697–706.

[3] M. Langer, C. J. König, P. Gebhard, and E. André, "Dear computer, teach me manners: Testing virtual employment interview training," *International Journal of Selection and Assessment*, vol. 24, no. 4, pp. 312–323, 2016.

[4] S. Samrose, R. Zhao, J. White, V. Li, L. Nova, Y. Lu, M. R. Ali, and M. E. Hoque, "Coco: Collaboration coach for understanding team dynamics during video conferencing," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 4, pp. 1–24, 2018.

[5] S. Samrose, D. McDuff, R. Sim, J. Suh, K. Rowan, J. Hernandez, S. Rintel, K. Moynihan, and M. Czerwinski, "Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[6] S. Samrose and E. Hoque, "MIA : Motivational Interviewing Agent for Improving Conversational Skills in Remote Group Discussions," *PACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–24, 2022.

[7] Y. R. Tausczik and J. W. Pennebaker, "Improving teamwork using real-time language feedback," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 459–468.

[8] M. I. Tanveer, E. Lin, and M. Hoque, "Rhema: A real-time in-situ intelligent interface to help people with public speaking," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 286–295.

[9] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE transactions on multimedia*, vol. 16, no. 4, pp. 1018–1031, 2014.

[10] S. Okada, L. S. Nguyen, O. Aran, and D. Gatica-Perez, "Modeling dyadic and group impressions with intermodal and interperson features," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–30, 2019.

[11] T. Onishi, A. Yamauchi, A. Ogushi, R. Ishii, A. Fukayama, T. Nakamura, and A. Miyata, "Modeling japanese praising behavior by analyzing audio and visual behaviors," *Frontiers in Computer Science*, vol. 4, p. 29, 2022.

[12] R. Li, J. Curhan, and M. Hoque, "Understanding social interpersonal interaction via synchronization templates of facial events," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[13] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. Junior, C. Jacques, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund *et al.*, "Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2177–2188.

[14] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, 2021.

[15] C. Li, W. Wang, B. Balducci, D. Marinova, and Y. Shang, "Predicting conversation outcomes using multimodal transformer," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–6.

[16] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[17] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[18] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 28, pp. 1849–1863, 2020.

[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[20] S. Rasipuram, J. H. Bhat, A. Maitra, B. Shaw, and S. Saha, "Multimodal depression detection using task-oriented transformer-based embedding," in *2022 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2022, pp. 01–04.

[21] S. Ghosh, G. V. Singh, A. Ekbal, and P. Bhattacharyya, "Comma-deer: Common-sense aware multimodal multitask approach for detection of emotion and emotional reasoning in conversations," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 6978–6990.

[22] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 839–848.

[23] P. Xu and X. Zhu, "Deepchange: A large long-term person re-identification benchmark with clothes change," *arXiv preprint arXiv:2105.14685*, 2021.

[24] Y.-S. Wang, H.-Y. Lee, and Y.-N. Chen, "Tree transformer: Integrating tree structures into self-attention," *arXiv preprint arXiv:1909.06639*, 2019.

[25] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, 2021.

[26] S. Pramanick, A. Roy, and V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3930–3940.

[27] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.

[30] N. Hojo, S. Kobashikawa, S. Mizuno, and R. Masumura, "Multimodal negotiation corpus with various subjective assessments for social-psychological outcome prediction from non-verbal cues," in *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC'22)*, 2022, pp. 6794–6801.

[31] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[32] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[33] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[34] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer, 2013, pp. 117–124.

[37] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[38] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] S. Okada, Y. Ohtake, Y. I. Nakano, Y. Hayashi, H.-H. Huang, Y. Takase, and K. Nitta, "Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 169–176.

[41] S. Dowlagar and R. Mamidi, "OFFLangOne@ DravidianLangTech-EACL2021: Transformers with the class balanced loss for offensive language identification in dravidian code-mixed text." in *Proceedings of the first workshop on speech and language technologies for dravidian languages*, 2021, pp. 154–159.

[42] S. Orihashi, M. Ihori, T. Tanaka, and R. Masumura, "Unsupervised domain adaptation for dialogue sequence labeling based on hierarchical adversarial training." in *Proceedings of INTERSPEECH*, 2020, pp. 1575–1579.