Real-time Emotion Pre-Recognition in Conversations with Contrastive Multi-modal Dialogue Pre-training

Xincheng Ju
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
xcju@stu.suda.edu.cn

Dong Zhang*
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
dzhang@suda.edu.cn

Suyang Zhu
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
syzhu@suda.edu.cn

Junhui Li
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
lijunhui@suda.edu.cn

Shoushan Li
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
lishoushan@suda.edu.cn

Guodong Zhou
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
gdzhou@suda.edu.cn

Code:None.

—— CIKM 2023

2023.11.23

Reported by Tingting Zhang

Leibniz-Institut für Sozialwissenschaften

Motivation

·SER (Static Emotion Recognition)

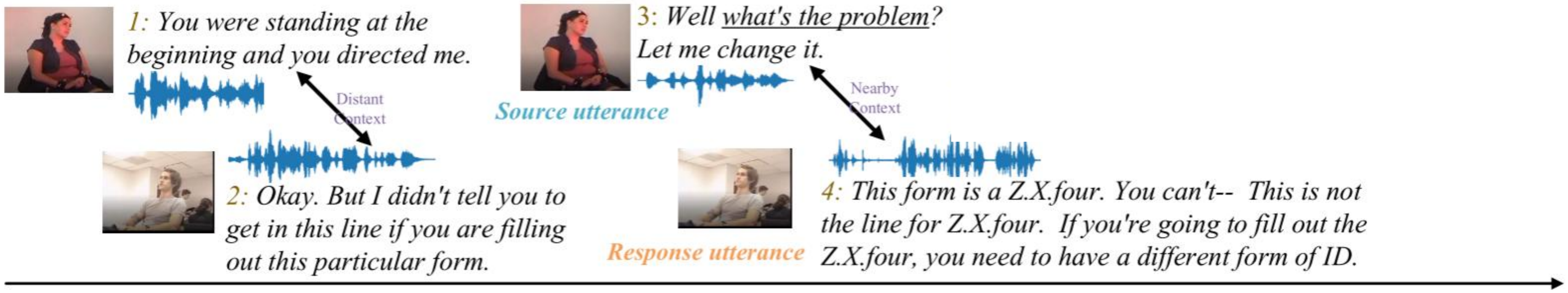
Many studies with focus on static emotion recognition, leverage past and future context to speculate the emotion of the existing target utterance in the textual or multi-modal scenario.

·RERC(Real-time Emotion Recognition in Conversations)

Recently, textual and multi-modal approaches observe the practical applications of RERC, which only leverage the past context to detect the target utterance emotion.

·MREPC (Multi-modal Real-time Emotion Pre-recognition in Conversations)

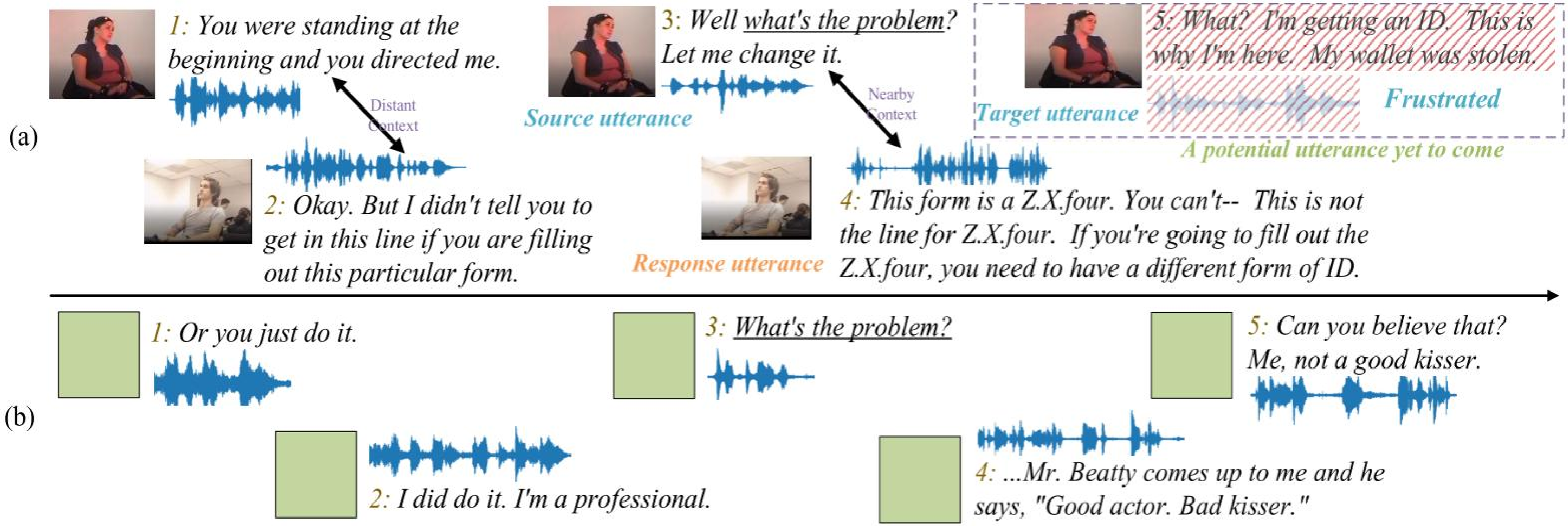
The objective is to predict the emotion of a forthcoming target utterance that is highly likely to occur(lack the target utterance and can only depend on the historical context).



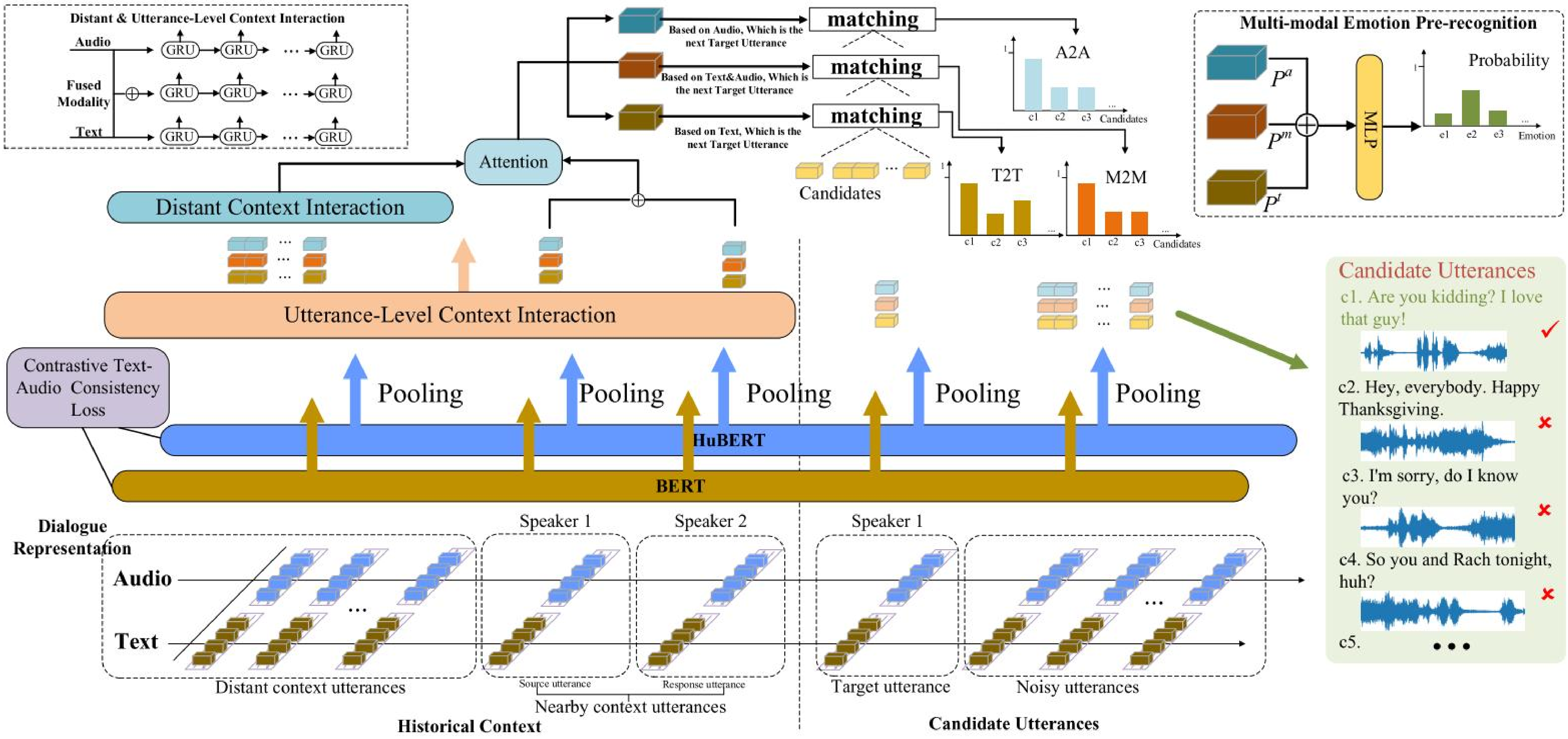
Motivation

It's difficult to speculate the emotion of the target utterance, due to the fact that the target utterance is not existing and the historical context can not provide adequate clues for emotion pre-recognition.

While, a complete conversation though unlabelled in Figure 1(b) may greatly supply the empathic information for (a).



Overview


Distant & Utterance-Level Context Interaction
Audio
Fused Modality
Text
GRU
Based on Audio, Which is the next Target Utterance
Based on Text&Audio, Which is the next Target Utterance
Based on Text, Which is the next Target Utterance
matching
A2A
T2T
M2M
Candidates
Multi-modal Emotion Pre-recognition
MLP
Probability
Emotion
Attention
Distant Context Interaction
Utterance-Level Context Interaction
Contrastive Text-Audio Consistency Loss
Pooling
HuBERT
BERT
Candidate Utterances
c1. Are you kidding? I love that guy!
c2. Hey, everybody. Happy Thanksgiving.
c3. I'm sorry, do I know you?
c4. So you and Rach tonight, huh?
c5.
Dialogue Representation
Speaker 1
Speaker 2
Distant context utterances
Source utterance
Response utterance
Nearby context utterances
Target utterance
Noisy utterances
Historical Context
Candidate Utterances

Method

1.Task Definition

$$p(y|\mathcal{D}) = p(e_{N+1}|C_d, C_n) \tag{1}$$

2.Modality Encoding

$$T_i^t = \mathsf{BERT}(u_i^t) \qquad i \in \{1, 2, \cdots, N\} \tag{2}$$

$$T_i^a = \mathsf{HuBERT}(u_i^a) \qquad i \in \{1, 2, \cdots, N\} \tag{3}$$

3.Utterance Representation

$$H_i^{\{t,a\}} = \mathsf{MaxPool}(T_i^{\{t,a\}}) + \mathsf{MeanPool}(T_i^{\{t,a\}}) \tag{4}$$

$$H_c, H_s, H_r = H_{\{1:N-2\}}, H_{N-1}, H_N \tag{5}$$

4.Utterance − level Context Interaction

$$H^m = W_m(H_i^t \oplus H_i^a) + b_m \quad i \in \{1, 2, \cdots, N\} \tag{6}$$

$$\hat{H}^\beta = \mathsf{GRU}^\beta(H^\beta) \tag{7}$$

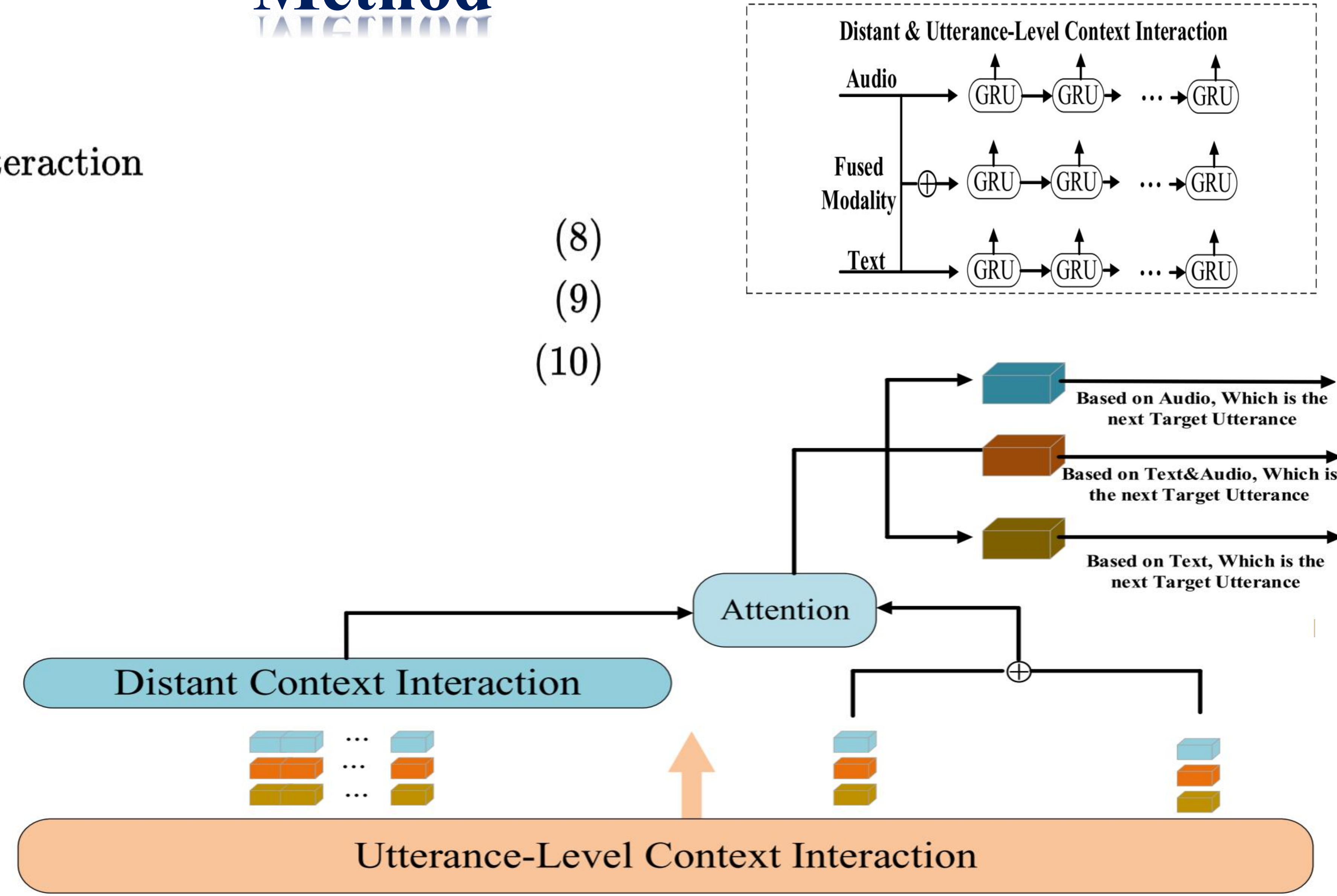
Method

5.Distant and Nearby Context Interaction

$$C^{\beta} = \mathsf{GRU}^{\beta}(\hat{H}_c^{\beta}) \tag{8}$$

$$Q^{\beta} = (H_s^{\beta} \oplus H_r^{\beta})W_{q\beta} + b_{q\beta} \tag{9}$$

$$P^{\beta} = \mathsf{Attention}(Q^{\beta}, C^{\beta}) \tag{10}$$



Method

1.Intra – modal Target Utterance Searching

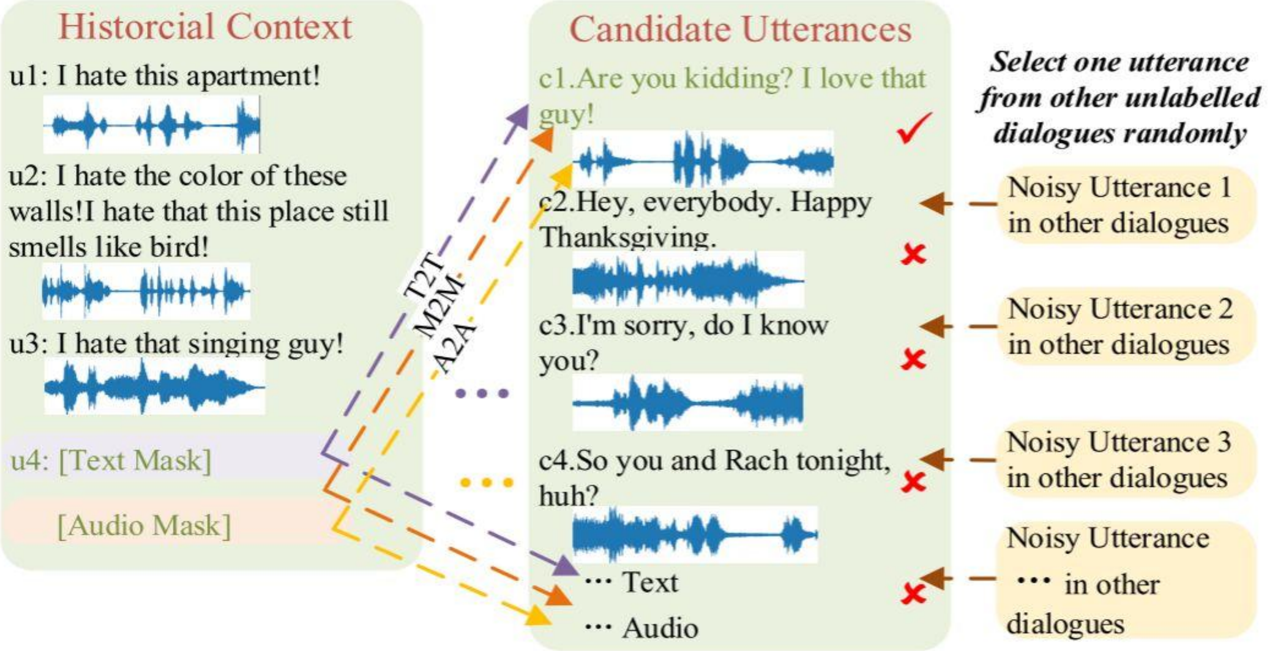
$$s_i^{t\to t} = \sigma((P^t)^\top K_i^t) \qquad i \in \{1,2,\cdots,k\} \tag{11}$$

$$s_i^{a\to a} = \sigma((P^a)^\top K_i^a) \qquad i \in \{1,2,\cdots,k\} \tag{12}$$

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

2.Inter – modal Target Utterance Searching

$$s_i^{m\to m} = \sigma((P^m)^\top K_i^m) \qquad i \in \{1,2,\cdots,k\} \tag{13}$$



Method

3.Contrastive Loss of Target Utterance Searching

$$\mathcal{L}_{Intra} = -\log(s_i^{t\to t} \cdot s_i^{a\to a}) + \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_i^{t\to t} \cdot s_i^{a\to a}) \tag{14}$$

$$\mathcal{L}_{Inter} = -\log(s_i^{m\to m}) + \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_i^{m\to m}) \tag{15}$$

$$\mathcal{L}_{II} = \mathcal{L}_{Intra} + \mathcal{L}_{Inter} \tag{16}$$

Method

4.Contrastive Loss of Text − Audio Consistency

$$logits = H^t \cdot (H^a)^\top \tag{17}$$

$$\mathcal{L}_h = -\sum_{i=1}^{N} \log\left(\frac{\exp^{logits[i,i]}}{\sum_{j=1}^{N} \exp^{logits[i,j]}}\right) \tag{18}$$

$$\mathcal{L}_v = -\sum_{i=1}^{N} \log\left(\frac{\exp^{logits[i,i]}}{\sum_{j=1}^{N} \exp^{logits[j,i]}}\right) \tag{19}$$

$$\mathcal{L}_{hv} = (\mathcal{L}_h + \mathcal{L}_v)/2 \tag{20}$$

5.Total Pre − training Loss

$$\mathcal{L}_{total} = \zeta\mathcal{L}_{II} + \eta\mathcal{L}_{hv} \tag{21}$$

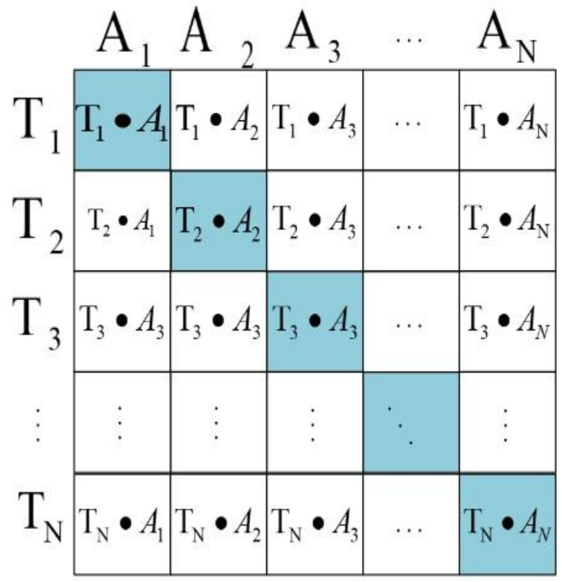


Figure 4: Contrastive Text-Audio Consistency Loss. T:Text, A:Audio

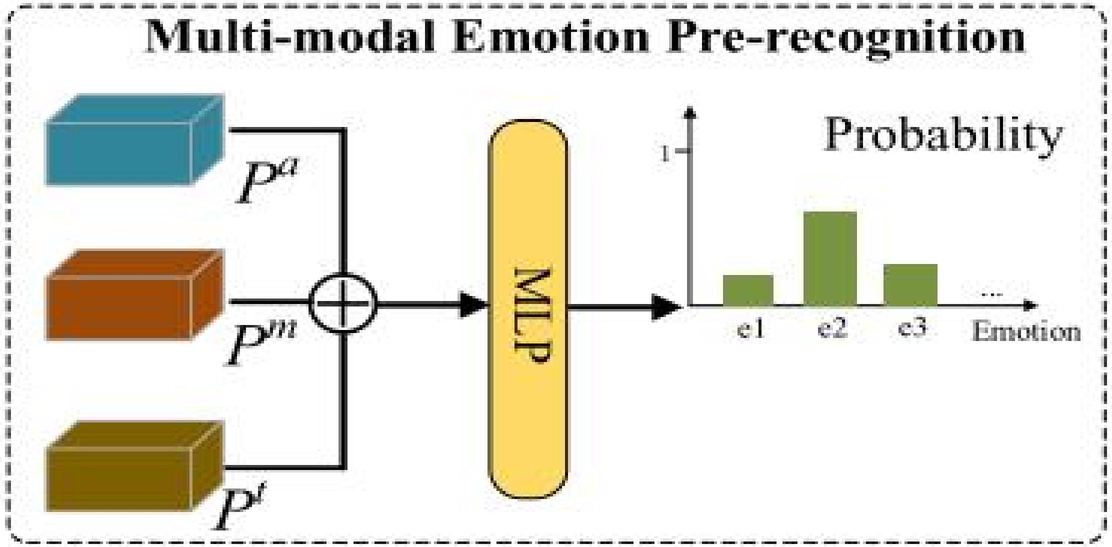
Method

1.Pre − recognition

$$\tilde{U} = \rho(\varrho((P^t \oplus P^m \oplus P^a)W_1 + b_1)W_2 + b_2)) \quad (22)$$

2.Loss of MREPC

$$\mathcal{J} = -\omega(y) \cdot log(\tilde{U}^y) \quad (23)$$



Experiments

$$\mathcal{L}_{Intra} = -\log(s_i^{t\to t} \cdot s_i^{a\to a}) + \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_i^{t\to t} \cdot s_i^{a\to a}) \quad (14)$$

$$\mathcal{L}_{Inter} = -\log(s_i^{m\to m}) + \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_i^{m\to m}) \quad (15)$$

$$\mathcal{L}_{II} = \mathcal{L}_{Intra} + \mathcal{L}_{Inter} \quad (16)$$

Approaches	TUS	$R_5@1$	$R_5@2$	$R_{11}@1$	$R_{11}@2$
Intra-modal	T2T	41.18	63.74	24.60	40.13
	A2A	96.38	97.20	96.89	96.93
Inter-modal	M2M	80.82	82.79	79.96	80.91
Multi-modal	T2T	45.36	67.12	27.88	41.70
	A2A	87.57	89.59	86.05	87.50
	M2M	90.59	92.84	89.33	90.46

Table 3: The performance of target utterance searching (TUS) in pre-training with different perspectives. The intra-modal part comes from our model which only calculates the intra-modal TUS loss (Eq. 14), while the inter-modal part only calculates the inter-modal TUS loss (Eq. 15). Multi-modal part denotes that both the intra- and inter-modal TUS loss are calculated simultaneously (Eq. 16). Note that higher scores mean better experimental results.

Experiments

Modality	Approaches	P-MELD WA	P-MELD AA	P-MELD F1	P-IEM WA	P-IEM AA	P-IEM F1
Text	PPE-Text	45.5	33.3	20.9	29.4	10.0	4.5
	NSF-Text-our impl.	48.6	41.3	35.6	45.2	17.9	12.0
	AGHMN-Text	43.4	36.4	32.1	41.7	20.7	19.4
	DialogXL-Text	45.5	33.3	20.9	44.9	20.5	18.8
Text+Audio	PPE[9]	43.3	33.9	29.2	34.8	12.6	9.3
	NSF-our impl. [39]	49.2	41.3	36.4	44.4	17.5	11.7
	AGHMN [15]	44.7	33.7	30.5	35.4	13.1	9.5
	DialogXL [35]	42.4	35.3	31.3	45.3	21.1	18.3
	DialogXL-our impl.	46.3	36.9	32.5	48.8	27.9	26.4
	BiDDIN [42]	42.2	34.4	30.5	34.3	14.7	12.7
	MMGCN [13]	46.5	34.6	24.7	36.9	22.8	20.5
	MDI-our impl. [44]	46.7	40.0	35.2	42.6	17.3	11.8
	TCMP(ours)	**50.1**	**41.5**	**38.9**	**53.9**	**28.2**	**27.4**

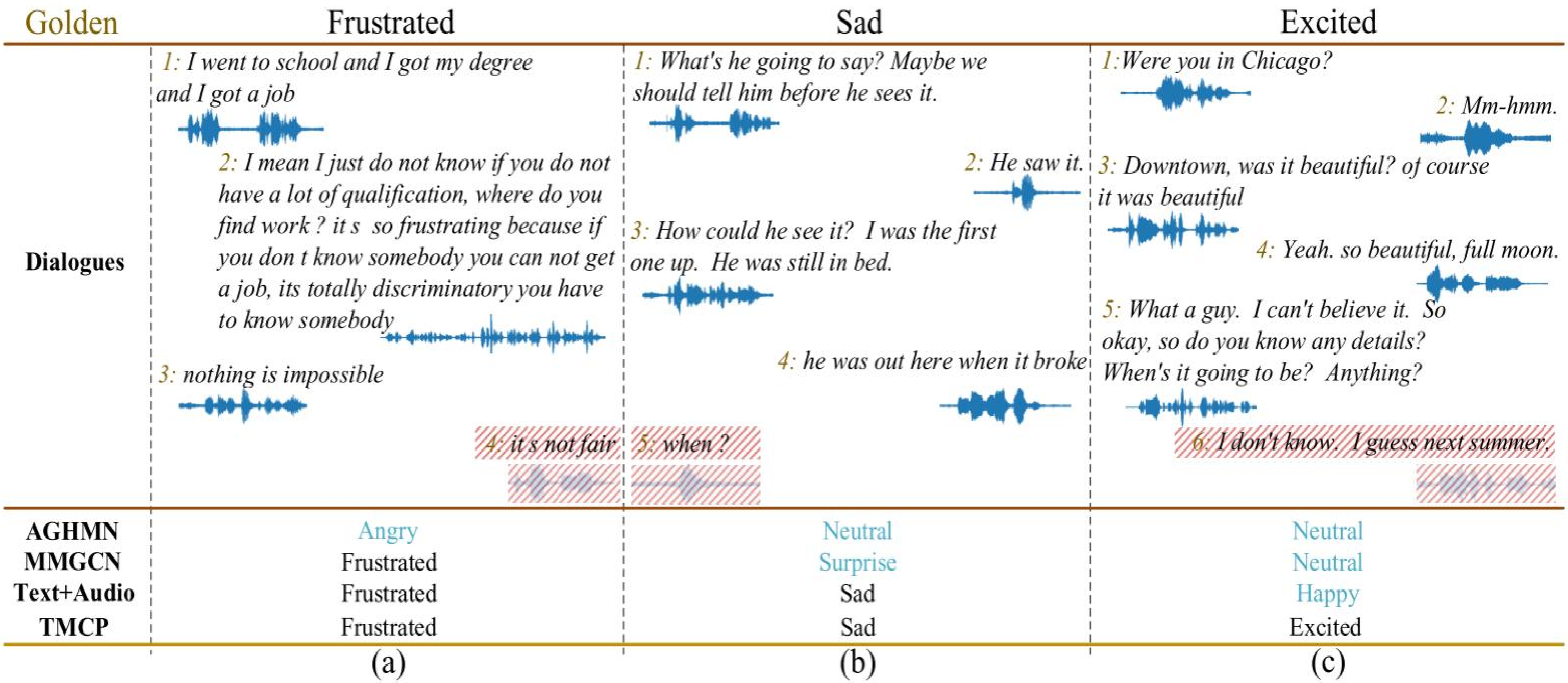
Table 4: The performance of different approaches for MREPC task. Text: only utilize textual modality. our impl.: implementing our modified setting for the corresponding approach.

Experiments

Approaches	P-MELD			P-IEM		
	WA	AA	F1	WA	AA	F1
Text w/o Dist	49.2	41.2	36.7	44.8	17.8	11.9
Text	49.5	41.7	36.9	44.6	17.6	11.8
Text-Pre	49.3	41.1	37.8	48.9	21.2	18.8
Text+Audio w/o Dist	48.4	41.5	37.1	46.0	19.6	15.8
Text+Audio	49.6	42.1	36.8	52.7	27.3	26.8
TCMP(ours)	**50.1**	41.5	**38.9**	**53.9**	**28.2**	**27.4**

Table 5: The performance of single-modal and multi-modal ablated approaches on both datasets.
-Pre denotes the pre-trained approach.
w/o Dist denotes the elimination of distant context.

Experiments



Thanks!