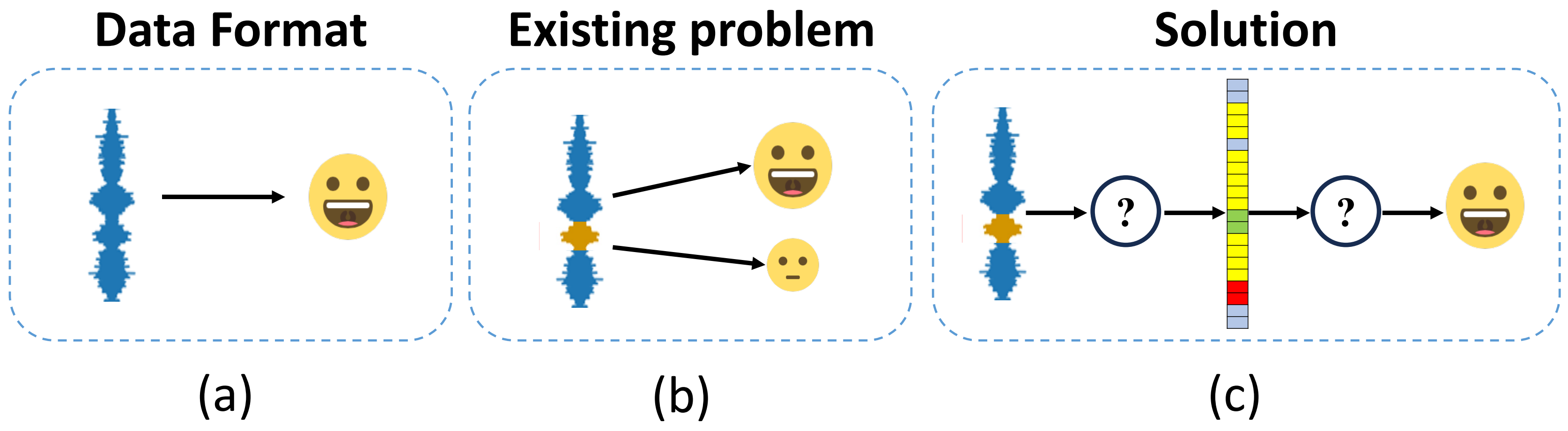
Frame-level Emotional State Alignment Method for Speech Emotion Recognition

Qifei Li, Yingming Gao, Cong Wang, Yayue Deng, Jinlong Xue, Yichen Han, Ya Li
Beijing University of Posts and Telecommunications, Beijing, China

ICASSP 2024 KOREA

Background

Motivation: **Not all frames** in an audio have **emotional states consistent** with **utterance-level** emotional **label**.



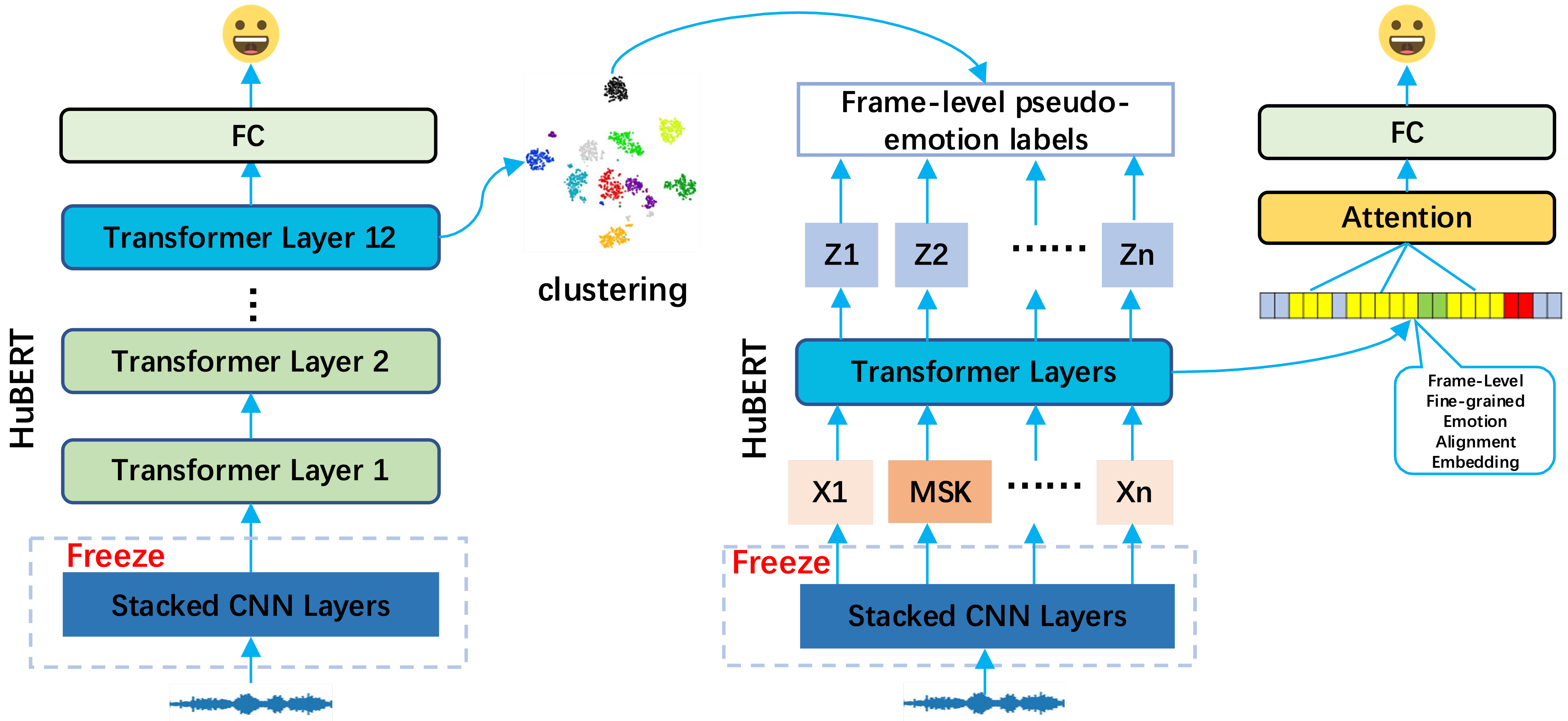
Contributions

□ Based on the previous research work, a frame-level **pseudo-emotional label generation** is proposed;

□ Proposing a method for **frame-level pseudo-emotion label and representation alignment** based on **masked language model**;

□ **Mapping frame-level** emotional alignment **representation**s and **utterance-level** emotional **labels** by **attention mechanisms** for SER, and **achieving remarkable performance**.

Methods



Phase 1. TAPT and cluster

Phase 2. Pre-training HuBERT for frame-level alignment

Phase 3. Fine-tuning CPT-HuBERT for SER

Phase 1:(Generate pseudo label for each frame)
□ Using task adaptive pretrain HuBERT with IEMOCAP.
□ Extracting the embedding of i-th transformer layer to generate pseudo emotion labels by K-means.

Phase 2:(Align pseudo label and representation)
□ Continuing pretraining HuBERT with frame-level pseudo-emotion labels to align frame-level pseudo label and representation of each frame.

Phase 3:(Realize SER)
□ Utilizing attention to map the frame-level emotion alignment representation and utterance-level label for SER.

Experiments and Results

Dataset:
- □ **IEMOCAP**, the number of samples is **5531**;
- □ **Leave-one-session-out** cross validation;
- □ Metrics are the **unweighted and weighted** accuracy.

Pretraining:
- □ **Masked language model (MLM)** for pretraining;
- □ **Mask 20%** representation along with time dimension;
- □ **Each fold of** CV requires pretraining a model.

Clustering:
- □ **K-means**;
- □ The number of clusters attempted is **50, 100, 150.**

Pooling:
- □ **Soft attention pooling；**
- □ **Average pooling.**

The ablation experiments of the numbers of cluster (Table 1):
- □ The best number of clusters for SER on IEMOCAP is **50**.
- □ The output of the **ninth** transformer layer in HuBERT-base for clustering is optimal.
- □ The **number** of clusters may **be related to** the **size** of the dataset.

The ablation experiments of the method of pooling (Table 1):
- □ Attention outperform average pooling in aligning frame-level representations and utterance-level labels;
- □ However, as the number of layers and clusters increases, the average pooling performance approaches and gradually surpasses the attention pooling.

Performance comparison with previous methods (Table 2):
- □ The proposed method **achieves optimal performance compared to other recent research results**.
- □ Meanwhile, the performance of method we proposed **approximates that of some multimodal methods**.

Table 1. The UA/WA (%) of ablation experiments of different poolings and different clusters of different transformer layers. BL means Baseline. The i-*th* layer represents extracting the embeddings from the i-*th* transformer layer to cluster.

Layers	Clusters	Average Pooling	Attention Pooling
BL [15]	-	74.3/-	-
TAPT	-	74.1/72.8	-
6-*th*	50	75.0/73.6	75.2/73.6
	100	74.8/73.3	75.1/73.5
	150	74.5/72.7	74.3/73.2
9-*th*	50	**75.1**/73.5	**75.7/74.7**
	100	75.0/**73.9**	75.3/74.0
	150	74.8/73.5	74.6/73.2
11-*th*	50	74.3/72.7	74.4/73.0
	100	74.0/72.8	74.2/72.7
	150	74.3/70.1	73.5/72.5

Table 2. Performance comparison of UA and WA with previous methods on IEMOCAP. The P-TAPT is baseline.

Type	Year	Methods	UA(%)	WA(%)
Audio	2023	P-TAPT [15]	74.3	-
	2023	SMW-CAT [20]	74.2	73.8
	2023	ShiftCNN [11]	74.8	72.8
	2023	SUPERB [13]	75.6	-
	-	**Ours**	**75.7**	**74.7**
Multi-modal	2023	MTG [21]	75.0	74.5
	2023	MSMSER [22]	76.4	75.2

Conclusions

□ **Proposing** an effective frame-level emotional state alignment **method** based on **MLM for SER** and achieving outstanding performance.

□ The performance **is strongly correlated with the representation** of the i-th **transformer** layer used **for clustering** and **the number of clusters**.

More Information

The QR of Paper

The QR of Code

Main References:

□ ***Exploring wav2vec2.0 fine-tuning for improved speech emotion recognition, Chen et al., ICASSP 2023.***

E-mail ： **liqifei@bupt.edu.cn**