LEVERAGING SPEAKER AND LISTENER PERSONALITIES AND THEIR INTERACTIONS FOR SPEECH EMOTION RECOGNITION

Zicheng Yuan, Yuan Gao, Yahui Fu, Yizhou Zhang, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{yuan,gao,fu,yizhang,kawahara}@sap.ist.i.kyoto-u.ac.jp

ABSTRACT

In this paper, we explore the effect of speaker and listener personality traits and their interactions on speech emotion recognition (SER) in conversations. We design an attention-based SER model that captures the interaction between speaker and listener personalities, while using their personality traits as auxiliary inputs to enhance SER. Experiments on the PA-IEMOCAP dataset demonstrate that incorporating both speaker and listener personality traits yields substantial gains. Moreover, explicitly modeling their interactions brings further improvements, achieving a 15.0% relative improvement in unweighted accuracy over the baseline. Since the ground-truth personality information is often unavailable in practice, we also use predicted personality traits as input, which results in a 12.3% relative improvement over the baseline. These findings highlight that personality traits, particularly their interactive effects, serve as a valuable source of information for enhancing SER.

***Index Terms*—** Speech emotion recognition, Big-Five personality, attention-based fusion

1. INTRODUCTION

Speech emotion recognition (SER) aims to recognize the emotional state of a speaker from speech [1] and is a key technology for building natural and user-friendly human–computer interaction [2]. SER enables systems to generate empathic response, thereby improving user experience [3]. Recent studies have shown that incorporating auxiliary tasks, such as automatic speech recognition (ASR) [4] and gender recognition [5], can improve SER performance. These findings indicate that, in addition to acoustic cues, leveraging such auxiliary information can effectively enhance SER. Our social experience suggests that people with different personalities tend to express their emotions in different ways [6]. Also, it is often easier for us to interpret the emotions of familiar individuals than those of strangers [7, 8]. These suggest that personality information may also provide useful cues for emotion recognition.

Previous research has revealed a strong correlation between personality traits and emotional expression [9, 10, 11], while some have attempted to incorporate personality information into emotion recognition tasks and shown that personality can indeed improve model performance [12, 13, 14]. However, these works only considered the individual influence of the speaker's personality on their own emotional expression, while ignoring the role of the listener's personality. In real conversational scenarios, people often adjust their communication style when facing different listeners [15]. Therefore, we hypothesize that emotional expression may be influenced not only by the speaker's own personality but also by the listener's personality.

Beyond influencing emotion expression, the listener's personality may even transiently affect the speaker's own personality state, creating interaction effects between them. Psychological research has shown that although personality is relatively stable over the long term, it exhibits considerable short-term fluctuations depending on situational context [16, 17] and can be influenced during social interactions [18]. For example, a person tends to temporarily become more outgoing when interacting with an extroverted listener, but may show more restraint when facing an introverted listener. This important perspective has not yet been systematically studied in the field of SER.

Motivated by this, we propose to explicitly incorporate personality traits of both speaker and listener into SER and to explore the effect of their interaction on SER. We build a multi-task learning framework, in which multi-head attention (MHA) [19] modules are used both for interaction and for fusion: the interaction modules first model the influence between the speaker's and listener's personalities, and the fusion modules then integrate their personality representations with acoustic features for the downstream SER task. We conduct extensive experiments on the PA-IEMOCAP dataset[1][2] [20, 13], and the results show that (1) incorporating both the speaker's and listener's personalities significantly improves SER performance; (2) explicitly modeling their interactions yields further gains; and (3) even using predicted personalities as input leads to stable improvements on SER. These findings demonstrate that personality traits of speaker and listener, as well as their interactions during conversations, are an effective source of information for enhancing SER performance.

2. PROPOSED METHOD

To investigate the influence of personality traits and their interactions, we propose two modules: an interaction module between the speaker's and listener's personalities, and a fusion module that integrates personality traits into acoustic features (Fig. 1). We also build a personality recognition model to simulate scenarios where ground-truth personalities are unavailable.

2.1. Personality Interaction Module

We design a personality interaction module that models the interaction between the speaker's and the listener's personalities before fusing them into the acoustic features.

The Big-Five personality model consists of five dimensions (*Openness, Conscientiousness, Extraversion, Agreeableness, and*

[1] http://hdl.handle.net/2433/298481

[2] https://github.com/Kyoto-University-Speech-and-Audio/PA-IEMOCAP

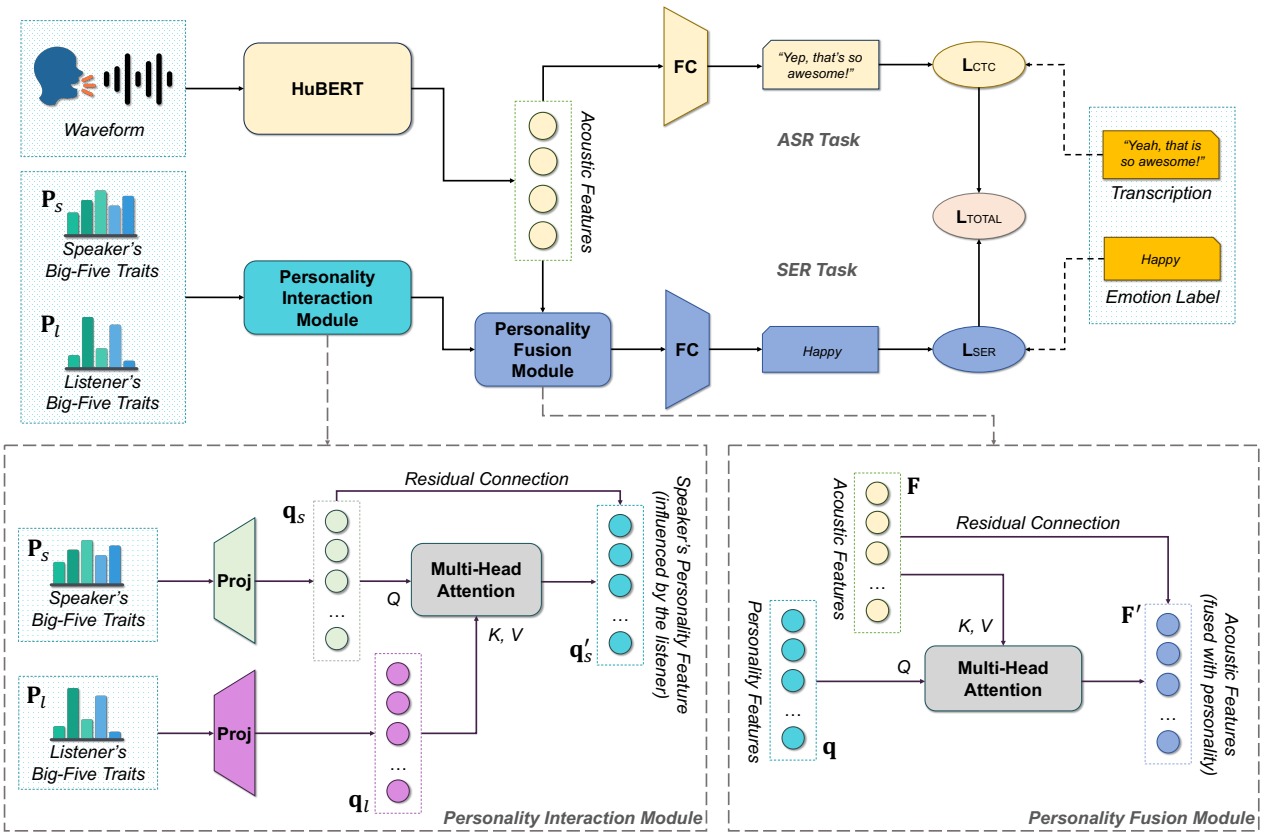

Fig. 1: Overall architecture of the SER model with the personality interaction module and the personality fusion module. The personality interaction module here illustrates the case where the listener's personality representation modulates the speaker's personality representation.

Neuroticism) [21], each rated on a 7-point scale (1–7). We concatenate the traits of the speaker and the listener into their personality vectors, denoted as $\mathbf{P}_s$ and $\mathbf{P}_l$, respectively. Let H denote the hidden dimensionality of the acoustic features. We then map these two personality vectors into the same H-dimensional space using separate linear projection layers W_s and W_l to obtain personality representations $\mathbf{q}_s$ and $\mathbf{q}_l$:

$$\mathbf{q}_s = W_s \mathbf{P}_s, \tag{1}$$

$$\mathbf{q}_l = W_l \mathbf{P}_l. \tag{2}$$

Then, we apply multi-head attention (MHA) [19] to model the interaction between personality representations. For example, when modeling how the listener's representation modulates the speaker's representation, we use the speaker's representation $\mathbf{q}_s$ as the query and the listener's representation $\mathbf{q}_l$ as the key and value. The output is then added back to $\mathbf{q}_s$ through a residual connection to obtain the updated speaker personality representation $\mathbf{q}'_s$ (Eq. 3). A similar procedure is applied for the opposite direction (Eq. 4).

$$\mathbf{q}'_s = \mathbf{q}_s + \mathrm{MHA}(query = \mathbf{q}_s,\ key = \mathbf{q}_l,\ value = \mathbf{q}_l), \tag{3}$$

$$\mathbf{q}'_l = \mathbf{q}_l + \mathrm{MHA}(query = \mathbf{q}_l,\ key = \mathbf{q}_s,\ value = \mathbf{q}_s). \tag{4}$$

The direction of interaction is controllable: it can be configured as listener to speaker, speaker to listener, or bidirectional. The lower-left dashed box in Fig. 1 shows the situation where the listener has influence on the speaker's personality. Through this design, our model can explicitly model the interaction between the speaker's and the listener's personalities.

2.2. Personality Fusion Module

The lower-right dashed box in Fig. 1 shows the structure of personality fusion module. For each utterance, we apply MHA mechanism that takes the personality representation $\mathbf{q}$ as the query and the frame-level acoustic features $\mathbf{F} \in \mathbb{R}^{T \times H}$ extracted by the HuBERT encoder as the key and value. The resulting context is added to the original acoustic features via a residual connection to get acoustic features fused with personality representation, denoted as $\mathbf{F}'$:

$$\mathbf{F}' = \mathbf{F} + \mathrm{MHA}(query = \mathbf{q}, key = \mathbf{F}, value = \mathbf{F}), \tag{5}$$

where $\mathbf{q}$ can be $\mathbf{q}'_s$, $\mathbf{q}'_l$, $\mathbf{q}_s$ or $\mathbf{q}_l$.

When both the speaker and listener personalities are used, we design two different fusion strategies:

(1) Sequential. We first fuse the speaker's personality and then the listener's personality through two successive steps:

$$\begin{aligned} \mathbf{F}_s &= \mathbf{F} + \mathrm{MHA}(query = \mathbf{q}'_s,\ key = \mathbf{F},\ value = \mathbf{F}), \\ \mathbf{F}' &= \mathbf{F}_s + \mathrm{MHA}(query = \mathbf{q}'_l,\ key = \mathbf{F}_s,\ value = \mathbf{F}_s). \end{aligned} \tag{6}$$

(2) Parallel. We stack the two personalities together, and let them be jointly fused to the acoustic features via a single MHA. The two attended outputs are then averaged and added back to the acoustic features:

$$\begin{aligned} \mathbf{Q} &= \mathrm{Concat}(\mathbf{q}'_s, \mathbf{q}'_l) \in \mathbb{R}^{2 \times H}, \\ \mathbf{F}' &= \mathbf{F} + \overline{\mathrm{MHA}}(query = \mathbf{Q},\ key = \mathbf{F},\ value = \mathbf{F}). \end{aligned} \tag{7}$$

Finally, $\mathbf{F}'$ is used in the downstream SER task.

2.3. Speech Emotion Recognition Model

Our SER model is built upon a multi-task learning framework, which is extended with two additional modules introduced in Section 2.1 and Section 2.2. The upper part of Fig. 1 shows the main structure of this model.

The backbone model is based on a HuBERT-base encoder that produces frame-level acoustic representations, with an ASR task trained jointly via the connectionist temporal classification (CTC) loss [22]. The utterance-level representations are then used for emotion classification. The overall training objective is a weighted sum of the two losses:

$$\mathcal{L} = \mathcal{L}_{\text{SER}} + \alpha \mathcal{L}_{\text{CTC}}, \tag{8}$$

where α is the weighting factor for the CTC loss.

2.4. Personality Recognition Model

To obtain personality information from speech, we train a multi-task personality recognition (PR) model similar to the SER model described in Section 2.3. Given a conversation consisting of multiple utterances from the same speaker, we first extract frame-level acoustic features from each utterance and apply mean pooling over the frames to obtain utterance-level representations. These representations are stacked and averaged to produce a conversation-level embedding, which is then fed into a linear regression head to predict a scalar score for one personality trait. We train separate models for each of the Big-Five traits (*Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*), and then concatenate the outputs to form final results. Similarly, we have the training objective:

$$\mathcal{L} = \mathcal{L}_{\text{PR}} + \beta \mathcal{L}_{\text{CTC}}, \tag{9}$$

where β is the weighting factor for the CTC loss.

The predicted personality traits are used as input to the SER model described in Section 2.3 to evaluate their effectiveness, thereby simulating a real application scenario where it is difficult to obtain ground-truth personality traits.

Note that reliable personality prediction requires a considerable amount of speech utterances [13], as even humans cannot judge personality from only a few utterances. Thus, we provide the model with a whole conversation when training and inference.

3. EXPERIMENTS

3.1. Data Preparation

We use the PA-IEMOCAP dataset [13], an extension of IEMOCAP [20] with Big-Five personality annotations for each speaker. It contains five sessions (01–05), and each session is divided into two parts. Each part is either male (M) or female (F), depending on the gender of the person who is recorded with motion capture in it.

For each session, the F part is used for validation, the M part for testing, and the other four sessions for training. This setting helps minimize potential information leakage between the training, validation, and testing sets.

We remove non-lexical special tokens in the transcriptions and discard utterances that contain no lexical content. For SER, we further filter out samples annotated with rare emotion labels, retaining only those labeled with the four target emotions (*angry, happy, sad, neutral*). As a result, the number of samples decreases from 10,039 to 9,980 after removing non-lexical utterances, and further to 5,500 after selecting only the four target emotions.

All experiment results are averaged over the five splits.

Table 1: Personality recognition results. MSE: Mean Squared Error; CCC: Concordance Correlation Coefficient.

Trait	CCC	MSE
Openness	0.703	0.579
Conscientiousness	0.652	0.682
Extraversion	0.758	0.470
Agreeableness	0.830	0.927
Neuroticism	0.749	0.987
Average	0.738	0.729

3.2. Implementation

All models were implemented in PyTorch [23] using the HuggingFace Transformers library [24], with HuBERT-base as the acoustic encoder. We used a batch size of 8, a learning rate of 5×10^{-5}, and 8 attention heads for all MHA layers, and optimized the models using AdamW [25] (β_1=0.9, β_2=0.999, ϵ=1e$-$8, weight decay 0.01) with a linear learning rate scheduler and a warmup ratio of 0.1.

For SER task, models were trained for maximum 100 epochs with ASR loss weight $\alpha = 0.1$. Early stopping was applied: training was stopped if the unweighted accuracy (UA) on validation set did not improve for 20 consecutive epochs, and the checkpoint with the highest validation UA was saved. For PR task, we adopted a two-stage training schedule: an ASR-only warm-up for 50 epochs starting from HuBERT-base with the first six encoder layers frozen, and then full training for another 100 epochs with ASR loss weight $\beta = 0.1$. The checkpoint with the highest concordance correlation coefficient (CCC) on validation set was saved.

4. RESULTS AND ANALYSIS

4.1. Personality Recognition Results

Table 1 summarizes the results of personality prediction. Our 5-fold experiments with held-out test sets achieved an average CCC of 0.738. Although this is slightly lower than the 0.778 reported in the previous work [13], our results show a similar trend: *Agreeableness* achieved the highest CCC, while *Conscientiousness* remained the most challenging to predict.

The predicted personalities were then used in the subsequent SER experiments, where we can also see performance gap from the ground-truth (GT) (Table 2), indicating that our predicted personalities are sufficiently reliable and informative.

4.2. SER Results without Personality Interaction

Table 2 shows the results of SER experiments under various configurations. We first compare the performance of integrating either the speaker's or the listener's personality traits into acoustic features without modeling personality interactions (No. 1&2). Both ground-truth and predicted personalities bring clear improvements over the baseline (No. 0). Even using only the listener's personality (No. 2) yields performance gains, although smaller than using the speaker's personality (No. 1). This indicates that the listener's personality contains useful prior information related to the speaker's emotion expression, helping the model better interpret emotional variations.

Moreover, combining both the speaker's and the listener's personality traits (No. 3&4) further improves performance. This suggests that the model benefits not merely from more information, but

Table 2: SER results using ground-truth (GT) and predicted (Pred) personality traits. "speaker" and "listener" show the personality features used for speaker and listener respectively. $\mathbf{q}_s$/$\mathbf{q}_l$: original personality traits; $\mathbf{q}'_s$/$\mathbf{q}'_l$: personality traits after interaction. "fusion" shows the fusion method. "Rel." denotes the relative improvement over the baseline (No. 0).

No.	Configuration			GT				Pred			
	speaker	listener	fusion	UA	ACC	UA Rel.	ACC Rel.	UA	ACC	UA Rel.	ACC Rel.
0	-	-	-	0.7154	0.7138	-	-	0.7154	0.7138	-	-
1	$\mathbf{q}_s$	-	-	0.7920	0.7869	10.71%	10.24%	0.7615	0.7573	6.44%	6.09%
2	-	$\mathbf{q}_l$	-	0.7752	0.7626	8.36%	6.84%	0.7572	0.7464	5.84%	4.57%
3	$\mathbf{q}_s$	$\mathbf{q}_l$	par	0.7947	0.7813	11.08%	9.46%	0.7685	0.7549	7.42%	5.76%
4	$\mathbf{q}_s$	$\mathbf{q}_l$	seq	0.8111	0.8050	13.38%	12.78%	0.7828	0.7712	9.42%	8.04%
5	$\mathbf{q}'_s$	-	-	0.7838	0.7764	9.56%	8.77%	0.7559	0.7487	5.66%	4.89%
6	-	$\mathbf{q}'_l$	-	0.7834	0.7757	9.51%	8.67%	0.7610	0.7527	6.37%	5.45%
7	$\mathbf{q}'_s$	$\mathbf{q}_l$	par	0.8101	0.8064	13.24%	12.97%	**0.8037**	0.7978	**12.34%**	11.77%
8	$\mathbf{q}'_s$	$\mathbf{q}_l$	seq	**0.8227**	0.8147	**15.00%**	14.14%	0.7822	0.7733	9.34%	8.34%
9	$\mathbf{q}_s$	$\mathbf{q}'_l$	par	0.8176	0.8088	14.29%	13.31%	0.7906	0.7801	10.51%	9.29%
10	$\mathbf{q}_s$	$\mathbf{q}'_l$	seq	0.8125	0.8048	13.57%	12.75%	0.7891	0.7774	10.30%	8.91%
11	$\mathbf{q}'_s$	$\mathbf{q}'_l$	par	0.7980	0.7896	11.55%	10.62%	0.7748	0.7633	8.30%	6.93%
12	$\mathbf{q}'_s$	$\mathbf{q}'_l$	seq	0.8110	0.8050	13.36%	12.78%	0.7853	0.7760	9.77%	8.71%

from learning patterns based on the relative differences and compatibility between the two personalities.

In terms of fusion strategies, the sequential method (No. 4) outperforms parallel (No. 3). We attribute this to the fact that in the parallel strategy, both personality representations are treated equally as queries to the attention layer, making it hard for the model to distinguish their roles. By contrast, the sequential method uses the speaker and then the listener in order, which provides explicit role information and may allow the model to implicitly capture interaction effects between the two personalities even without an explicit interaction module.

4.3. SER Results with Personality Interaction

When only one side's personality is fused (No. 5&6), introducing interaction degrades performance compared to their no-interaction counterparts (No. 1&2). We attribute this to semantic drift: the interacted personality feature deviates from its original meaning, but the model treats it as the speaker's true personality, effectively injecting noise that disrupts SER.

In contrast, when both the speaker's and the listener's personalities are fused (No. 7–12) and personality interaction is enabled, the performances of most settings further improved from the no-interaction setting (No. 3&4). This confirms that the model can indeed benefit from personality interactions. Notably, the *listener→speaker* interaction achieves the highest UA (No. 8 with GT & No. 7 with Pred), which aligns with our hypothesis that the listener's personality influences the speaker's personality state and emotional expression. Interestingly, the *speaker→listener* interaction (No. 9) also achieves the second-best performance in both settings, which suggests that the model may have learned to infer "how the speaker perceives the listener's personality," thereby helping it predict how the speaker would adjust their emotional expression toward different types of listeners.

4.4. SER Results with Bidirectional Interaction

When incorporating bidirectional personality interaction (No. 11&12), the performance does not noticeably surpass that of the no-interaction counterparts (No. 3&4). We attribute this to: (1) the PA-IEMOCAP dataset is relatively small, containing only 151 conversations, and bidirectional interaction introduces more parameters and dependencies that may cause overfitting; and (2) predicted personalities inevitably contain noise, and bidirectional interaction amplifies this noise through repeated mutual updates. Under such limited data conditions, unidirectional interaction provides a more stable inductive bias and is more likely to generalize well.

5. CONCLUSIONS

In this study, we have investigated the effect of both speaker and listener personality traits, as well as their interactions, on speech emotion recognition tasks.

To this end, we introduced two attention-based modules: a personality interaction module, which captures the influences between speaker and listener personalities, and a personality fusion module, which integrates personality traits with acoustic features. Experiments on the PA-IEMOCAP dataset demonstrate that incorporating both speaker and listener traits substantially enhances SER performance, and explicitly modeling their interactions yields additional improvements. Furthermore, using predicted personality traits also results in consistent performance gains, demonstrating their effectiveness in practical applications.

These results highlight the importance of personality information, both individual and interactive, as a complementary resource for SER, and suggest new directions for developing socially aware emotion recognition systems.

6. ACKNOWLEDGMENT

This study was supported by the JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJMS2011) and JSPS KAKENHI 25H01142.

7. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Scott Brave and Cliff Nass, "Emotion in human-computer interaction," in *The human-computer interaction handbook*, pp. 103–118. CRC Press, 2007.

[3] Yahui Fu, Chenhui Chu, and Tatsuya Kawahara, "StyEmp: Stylizing empathetic response generation via multi-grained prefix encoder and personality reinforcement," in *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2024, pp. 172–185.

[4] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, "Speech emotion recognition with multi-task learning," in *Interspeech*. ISCA, 2021, vol. 2021, pp. 4508–4512.

[5] Yuan Gao, Hao Shi, Chenhui Chu, and Tatsuya Kawahara, "Enhancing two-stage finetuning for speech emotion recognition using adapters," in *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11316–11320.

[6] Lyndon A Mitchell, "The relationship between emotional recognition and personality traits," M.S. thesis, Rochester Institute of Technology, 2006.

[7] M Ida Gobbini, Ellen Leibenluft, Neil Santiago, and James V Haxby, "Social and emotional attachment in the neural representation of faces," *Neuroimage*, vol. 22, no. 4, pp. 1628–1635, 2004.

[8] Evy van Berlo, Thomas Bionda, and Mariska E Kret, "Attention toward emotions is modulated by familiarity with the expressor: A comparison between bonobos and humans," *Emotion*, vol. 23, no. 7, pp. 1904, 2023.

[9] Carroll E Izard, Deborah Z Libero, Priscilla Putnam, and O Maurice Haynes, "Stability of emotion experiences and their relations to traits of personality," *Journal of personality and social psychology*, vol. 64, no. 5, pp. 847, 1993.

[10] Yuanchao Li, Peter Bell, and Catherine Lai, "Transfer learning for personality perception via speech emotion recognition," in *Interspeech 2023*. ISCA, 2023, pp. 5197–5201.

[11] Md Ali Akber, Tahira Ferdousi, Rasel Ahmed, Risha Asfara, Raqeebir Rab, and Umme Zakia, "Personality and emotion—a comprehensive analysis using contextual text embeddings," *Natural Language Processing Journal*, vol. 9, pp. 100105, 2024.

[12] Le Zhang, Songyou Peng, and Stefan Winkler, "PersEmoN: A deep network for joint analysis of apparent personality, emotion and their relationship," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 298–305, 2019.

[13] Yuan Gao, Hao Shi, Yahui Fu, Chenhui Chu, and Tatsuya Kawahara, "Bridging speech emotion recognition and personality: Dataset and temporal interaction condition network," *IEEE Transactions on Affective Computing*, 2025.

[14] Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou, "Emotion recognition in conversation via dynamic personality," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 5711–5722.

[15] Howard Giles and Tania Ogay, "Communication accommodation theory," in *Explaining communication*, pp. 325–344. Routledge, 2013.

[16] William Fleeson, "Situation-based contingencies underlying trait-content manifestation in behavior," *Journal of personality*, vol. 75, no. 4, pp. 825–862, 2007.

[17] William Fleeson and Patrick Gallagher, "The implications of big five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis," *Journal of personality and social psychology*, vol. 97, no. 6, pp. 1097, 2009.

[18] Cornelia Wrzus and Brent W Roberts, "Processes of personality development in adulthood: The tessera framework," *Personality and Social Psychology Review*, vol. 21, no. 3, pp. 253–277, 2017.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] Lewis R Goldberg, "An alternative "description of personality": The big-five factor structure," in *Personality and personality disorders*, pp. 34–47. Routledge, 2013.

[22] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[25] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.