# Attention Learning with Retrievable Acoustic Embedding of Personality for Emotion Recognition

Jeng-Lin Li
*Electrical Engineering*
*National Tsing Hua University*
Taiwan
cllee@gapp.nthu.edu.tw

Chi-Chun Lee
*Electrical Engineering*
*National Tsing Hua University*
Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Modeling multimodal behavior streams to automatically identify emotion states of an individual has progressed extensively especially with the advancement of deep learning algorithms. Emotion, being an abstract internal state, creates substantial differences in an individual's behavior expressivity, the development of *personalized* recognition framework is a critical next step to improve algorithm's modeling capacity. In this work, we propose to integrate the target speaker's personality embedding into the learning of multimodal (speech and language) attention based network architecture to improve recognition performances. Specifically, we propose a Personal Attribute-Aware Attention Network (PAaAN) that learns its multimodal attention weights jointly with the target speaker's retrievable acoustic embedding of personality. Our acoustic domain adapted personality retrieval strategy mitigates the common issue on the lack of personality scores in the current available emotion databases, and our proposed PAaAN then learns its attention weight by jointly considering an individual target speaker's personality profile with his or her multimodal acoustic and lexical modalties. In this work, we achieve a 70% unweighted accuracy in the IEMOCAP 4-class multimodal emotion recognition task. Further analysis shows the effect of integrating personality on the variation of our attention weights of each acoustic and lexical behavior modality for each speaker in the IEMOCAP database.**

***Index Terms*—personality, multimodal emotion recognition, cross corpus retrieval, attention learning**

## I. Introduction

Recently, technological advancements have progressed tremendously to help the proliferation of a variety of human-centered applications, such as personalized advertisements, entertainment recommendations, and shopping experiences optimizations [1]–[3]. These user-centric applications require further computational advancement in deriving analytics to comprehensively understand an individual subject's internal states and traits, such as emotion and personality, in order to help personalize these services to achieve the next generation user experiences. In fact, computational frameworks based on machine learning and signal processing techniques have been proposed extensively to automatically recognize emotion states of an individual from measurable behavior data. Specifically, speech and language modalities, being the most natural human communication medium, have attracted substantial attention in the technical development effort of learning to recognize emotion from behavior signals (e.g., [4]–[6]).

Recent deep learning based approaches that jointly models both speech and language paralinguistics cues have received the state-of-the-art emotion recognition performances. Several notable multimodal speech and language emotion recognition works include: Sahay et al. introduce a relational tensor network that generates rich representations capturing the interaction of modalities to be used in recognizing emotion [7]. Lee et al. present a convolution attention to capture complex nonlinear correlation in the hidden multimodal feature space for emotion recognition [8]. Further, Cho et al. emphasize the design of different network architectures for acoustic and linguistic modality separately and perform late fusion with another layer of deep neural network [9]. While these works have achieved promising accuracy, they focus mainly on optimizing the frameworks with respect to the given emotion labels only, i.e., without considering that an individual's personal attributes such as personality would intricately impact the emotional expressive behaviors of each person at an *individual* level.

Personalized attributes of an individual can be seen as latent stable modulating factors dynamically influencing expressive behaviors of a person, which in terms create an individual differences in one's own emotion expressivity. For example, age, gender and culture are examples of personalized factors affecting one's belief, motives, and even stereotypical perceptions, which is then further realized in shaping an individual's thoughts, feelings and hence behaviors [10], [11]. Personality is one of the most important *personalized* attributes at an individual level underlie the variations in our actions, emotion expressions/regulations, and attitudes toward others [12]. In fact, the individual differences resulting from personal attributes of gender and culture are reported to be mediated by personality attributes [13], [14].

Individual personality is often assessed using a Big-Five inventory model, which includes dimensions of extraversion, agreeableness, openness, conscientiousness, and neuroticism [15]. Several of these attributes have been shown to be strongly tied to emotion, e.g., extroversion associating with outgoing and sociable person are inclined to be more positive, while neuroticism tendency is highly correlated to negative emotion [16]. Moreover, personality also relates to affect cognitive appraisal process and motivational structures, i.e., emotion regulation strategy and complicated emphatic emotion

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

processing for an individual under different contexts [13], [17]. Personality and emotion, while naturally affecting each other, only a handful of research has incorporated information of an individual's personality to enhance emotion recognition. For example, PersEmoN is one such framework proposed by Zhang et al. It is based on a deep multitask network structure to analyze the relationship between personality and emotion on facial expressions [18] and further extends to a personality-aware hyperplane construction method to improve emotion recognition within a multi-label prediction scheme [19]. Sagha et al. present a speech based framework by utilizing personality as a measure to identity different subgroups to perform valence-only recognition [20].

Most of these prior works either focus on single modality modeling or simply treating personality attribute as an auxiliary input. Without the joint modeling between behavior expressions and personality characteristics, it could limit the recognition modeling capacity. Furthermore, often these works assume the availability of Big-Five attributes score for each individual, which further limits the scalability of the algorithm in real world applications. In this work, we propose to integrate an acoustic-based personality embedding retrieval method to the Personalized Attribute-Aware Attention Network (PAaAN) to perform emotion recognition using acoustic and lexical modalities. PAaAN is an attention-based multimodal network architecture that integrates the target speaker's personality attributes into the learning of its multimodal (acoustic and lexical) attention mechanism. Since most emotion databases lack personality annotations, by leveraging other databases with personality scores and further utilize a cross corpus acoustic domain adapted retrieval strategy, we can represent the target speaker's personality as a retrievable embedding vector, i.e., computed as statistical functions of the retrieved personality scores.

We evaluate our proposed framework on the benchmark emotion database, the IEMOCAP database [21] using the SSPNET as our personality retrieval database [22]. We obtain a state-of-the-art 70% unweighted average recall (UAR) in 4-class emotion recognition task in the IEMOCAP, which improves 5.11% relative over the multimodal attention framework without retrievable acoustic embedding of personality. Moreover, we further analyze the learned PAaAN attention to explore the effect of personality on both audio and text modalities. The rest of paper is organized as follows: section 2 describes about the multimodal database and features; section 3 details our experimental setup, results, and analyses; finally, section 4 concludes with future works.

## II. Methods

### A. Databases

*1) Emotion Database - The IEMOCAP:* In this work, we evaluate our emotion recognition accuracy using a benchmark dyadic interaction database, the IEMOCAP [21]. It contains audio, video and word-aligned manual transcripts with an approximately 12-hour of data consists of 10039 utterances. The database includes 10 speakers paired into 5 dyad sessions. They perform both scripted and improvised sessions. In each session, the emotion labels are annotated by 3 raters at the utterance level. In this work, our 4-class emotion classification task is conducted on 5531 utterances with 1103 utterances for angry, 1636 for happy (includes excitement), 1084 for sad, and 1708 for neutral.

*2) Personality Database - The SSPNET:* We utilize a personality corpus consists of speech recording used in the Personality Sub-Challenge in Interspeech 2012 [22]. There are approximately a total of an hour and 40 minutes audio recordings comprise of 640 clips from 322 speakers in French news bulletins of Radio Suisse Romande. Each clip is rated by 11 raters based on Big-Five personality inventory including dimensions of openness, conscientiousness, extroversion, agreeableness, and neuroticism traits.

### B. Acoustic and Textual Representations

*1) Acoustic Features:* The acoustic features used for the emotion recognition are based on 45 dimensional low level descriptors (LLDs) including 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F0), loudness, voice probability, zero cross rate along with their first derivatives and the second derivatives of MFCCs and loudness. This LLDs set is extracted using the openSMILE toolbox [23]. The frame size is set as 60ms and the step size is 10ms. All the features are z-normalized speaker-wise, and the input time step to our PAaAN, which uses bi-directional long short term memory (BLSTM) network at core, is the average value of every 40ms.

On the other hand, we extract a 1583 dimensional acoustic features by computing a variety of statistical functions from the same set of 45 LLD features. This 1583 dimensional feature vector is termed as the Emobase 2010 in the openSMILE config file that is used in the personality embedding retrieval process in this work.

*2) Word Embeddings:* In this work, each word in transcripts is encoded using a pre-trained embedding GloVe originally trained with 42 billion tokens and 1.9 million vocabularies [24]. Given an utterance with $N$ words, an utterance can be encoded as a set of word embedding, $U = \{w_1, w_2, ..., w_N\}$ where $w$ is the word embedding and $U \in \mathbb{R}^{N\times 300}$. Thus, each word is encoded as a 300 dimensional vector as a time step for the PAaAN model.

### C. Multimodal Emotion Recognition Framework

The overall framework is demonstrated in Fig. 1, which contains three major components, i.e., acoustic domain adaptation, personality embedding retrieval, and Personal Attribute Aware Attention Network (PAaAN).

*1) Acoustic Domain Adaptation:* Since there is no annotated personality attributes in the IEMOCAP database, we derive our personality embedding retrieved from the SSPNet database. However, the SSPNet and the IEMOCAP are collected in a very different environment, we first implement a share-hidden-layer autoencoder (SHLA) approach to reduce the cross corpora acoustic domain discrepancy [20]. The

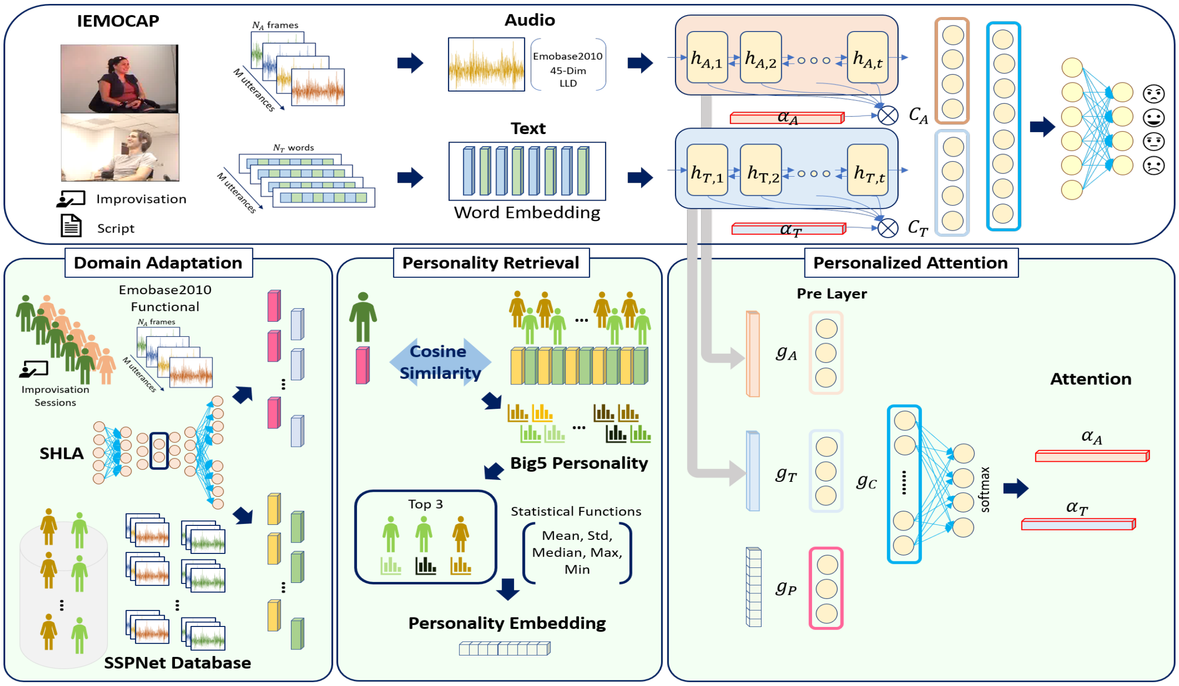

Fig. 1. *This is the overall PAaAN framework. We firstly perform acoustic domain adaption and utilize cosine similarity to retrieve top k similar utterances. We represent the target speaker's personality profile by computing statistics on these retrieved acoustic embedding of personality. Then, we introduce the personalized attention to jointly model inter-modality relation and personality embedding in a BLSTM framework for emotion recognition.*

SHLA network includes a shared parameters for both corpora in the encoder network, where the reconstruction networks are trained with corpus-specific layers. The shared encoder layers project both corpora data onto a common acoustic feature space while corpus-specific layers allow the variability of each corpus characteristics to be retained. In this work, we train the SHLA on the Emobase 2010 (1583 dimensions features) feature set and extract the latent layers for each corpus to perform personality embedding retrieval.

2) *Personality Embedding Retrieval:* The center idea of our retrieval approach is to represent the target speaker in the emotion corpus by making use of the available personality scores in the SSPNET personality corpus. After performing acoustic domain adaptation, we compute the cosine distance between an utterance in the emotion corpus and the audio clips in the personality database. We then retrieve the Big-Five personality scores from the top $k$ closest audio clips from the personality corpus. We collect all $k \times n$ retrieved personality ratings for a target subject in the emotion corpus, where $n$ is the total number of utterances spoken by the target subject. In order to derive a subject-level personality embedding for each speaker in the IEMOCAP database, we compute 5 statistical functions, i.e., mean, standard deviation, median, maximum and minimum, on these retrieved scores resulting in an embedding with 25 dimensions. This retrieval method can be conceptualized as representing *pseudo*-personality of the target speaker as measured across the utterances spoken in the IEMOCAP database from the perspective of the SSPNET personality corpus. Similar idea has recently been used in cross-corpus emotion perspective retrieval for data augmentation research [25], [26].

3) *Personal Attribute Aware Attention Network:* The overall architecture of PAaAN is composed of modality specific BLSTMs (BLSTM-A and BLSTM-T for audio and text respectively), where both are connected to a shared fully-connected emotion classification layers. The BLSTM hidden states of modality $m = \{T, A\}$ can be written as $h_{m,t} = [\overleftarrow{h_{m,t}} \bigoplus \overrightarrow{h_{m,t}}]$ where $\overleftarrow{h_{m,t}}$ and $\overrightarrow{h_{m,t}}$ denote the forward and backward hidden states. When learning the attention weights, we first use a single fully-connected layer as a pre-layer to condense information in $h_{A,t}$, $h_{T,t}$ and the derived subject-level personality embedding.

$$g_{m,t} = tanh(w_m^T h_{m,t} + b_m) \tag{1}$$

Then, these vectors ($g_{m,t}$) are concatenated together denoted as $g_{c,t}$, which is used to learn a personalized attention, $\alpha$. $\alpha$ is derived using time-normalized output of a fully-connected layer follow by a softmax function.

$$g_{c,t} = [g_{T,t}, g_{A,t}, g_{P,t}] \tag{2}$$

$$\alpha = \frac{\exp(g_{c,t})}{\sum_t^T \exp(g_{c,t})} \tag{3}$$

With this attention learning scheme, we learn separate acoustic and textual attention for BLSTM-A and BLSTM-T, i.e., $\alpha_A$ is learned specifically to re-weight acoustic low level descriptors while $\alpha_T$ is for re-weighting word embeddings. Although it is a modality specific re-weighting mechanism,

TABLE I

THE 4-CLASS RECOGNITION RESULTS IN THE EMOTION RECOGNITION TASK ARE DEMONSTRATED. THE LEFT PART SHOW THE MULTIMODAL BASELINES AND THE RIGHT PART PROVIDES COMPARISON AMONG MODELS WITH PERSONALITY EMBEDDING. DETAILED COMPARISON MODEL DESCRIPTIONS CAN BE FOUND IN SECTION III-A.

| | Audio | Text | A+T | Prev1 | Prev2 | I$_{A+T}$ | P-D | P-T | PAaAN$_{NA}$ | PAaAN$_{PP}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ang | 0.570 | 0.663 | 0.657 | 0.724 | 0.625 | 0.666 | 0.685 | 0.621 | 0.719 | **0.763** |
| hap | 0.527 | 0.647 | 0.666 | 0.675 | 0.652 | 0.685 | 0.751 | 0.598 | 0.669 | 0.730 |
| neu | 0.581 | 0.584 | 0.628 | 0.574 | 0.696 | 0.607 | 0.572 | 0.677 | 0.616 | 0.588 |
| sad | 0.608 | 0.551 | 0.666 | 0.665 | 0.633 | 0.705 | 0.693 | 0.601 | 0.712 | **0.720** |
| UAR | 0.571 | 0.611 | 0.654 | 0.659 | 0.651 | **0.666** | 0.675 | 0.624 | 0.679 | **0.700** |

the attention vectors are still learned by jointly modeling intermodality influences shown in equation 2. The information integration in attention enables a flexibility in determining the dynamics in the attention weighting to better enhance the emotion discriminative power. The time-wise outputs of BLSTM-A and BLSTM-T are multiplied with the personalized attention re-weighting vector $\alpha_A$ and $\alpha_T$, which gives rise to context vectors $C_A$ and $C_T$. The final representation is a concatenation of $C_A$ and $C_T$, and then it is fed into a fully connected network for final emotion classification.

## III. EXPERIMENTS

### A. Experimental Setup

In this work, we assess our model performance of the 4-class emotion recognition task on the benchmark multimodal emotion database IEMOCAP. The following is the list of comparison models:

- **Audio**: Baseline BLSTM attention model for recognizing emotion using acoustic features only
- **Text**: Baseline BLSTM attention model for recognizing emotion using word embedding only
- **A+T**: Baseline multimodal dual modality (speech and text) BLSTM attention model for recognizing emotion by concatenating $C_A$ and $C_T$
- **Prev1**: Comparison with the previous work using multimodal framework for audio and text on the same dataset [9]
- **Prev2**: Comparison with another previous work using multimodal framework for audio and text on the same dataset [27]
- **I$_{A+T}$**: Multimodal attention learning for recognizing emotion, i.e., PAaAN without personality embedding in the attention learning
- **P-D**: Including personality embedding by concatenating it with with $C_A$ and $C_T$
- **P-T**: Including personality embedding by duplicating the embedding to concatenate with the frame-level feature in the BLSTM training
- **PAaAN$_X$**: Using X approach to retrieve personality for PAaAN network

For $I_{A+T}$, it has the identical architecture as PAaAN but without integration of personality embedding; it is used as the baseline to examine the effectiveness of the personality integrated attention learning framework. Apart from PAaAN, two different embedding integration techniques, *P-D* and *P-T* are explored to investigate the ability of our framework in incorporating personality embedding into the recognition network II-C2. Our proposed PAaAN framework is denoted as $PAaAN_X$ where X can be either proposed embedding generating approach described in section II-C2 denoted as *PP* or the personality embedding derived by section II-C2 without domain adaption mentioned in II-C1 denoted as *NA*.

The following describes the details of our network parameters. The SHLA network has a symmetric auto-encoder structure of node size 1583-512-256-128-256-512-1583, and the128 dimensional latent feature vector is used for personality retrieval in section II-C2. We train the SHLA network with batch size 16 and learning rate 0.0001 for 10 epochs. For the PAaAN, we have the same structure for each modality-specific BLSTM, i.e., 128 nodes for both modalities, and the pre-layers $g_{m,t}$ are specified with 16 nodes. The concatenation of $C_A$ and $C_T$ is 128 nodes, and there is a 256-neuron fully connected layer followed by a softmax layer. The activation function for pre-layers are hyperbolic tangent whereas the other layers use relu. The model is trained jointly with 50 epochs, 32 batch size and 0.0001 learning rate. The experiments are conducted through leave-one-dyad-out cross validation, and we evaluate the results with unweighted averate recall (UAR).

### B. Results

Table I summarizes the IEMOCAP emotion classification results. Our proposed PAaAN with personality embedding, $PAaAN_{PP}$, outperforms all the other models. Specifically, it obtains 70% UAR that is 5.11% relative improvement compared to $I_{A+T}$, i.e., using the same framework without the integration of personality embedding. In terms of multimodal models, $I_{A+T}$ achieves 66.6%, which surpasses single modality model of audio or text only with 16.21% and 9.0% relative improvement respectively. Moreover, without personality embedding in the attention learning, it already exceeds performances of *A+T* and the other two recently-proposed multimodal frameworks [9], [27]. These results indicate the modeling capacity of our proposed attention learning for multimodal fusion, which provides a flexible yet powerful modeling approach to integrate speech and language modalities for emotion recognition.

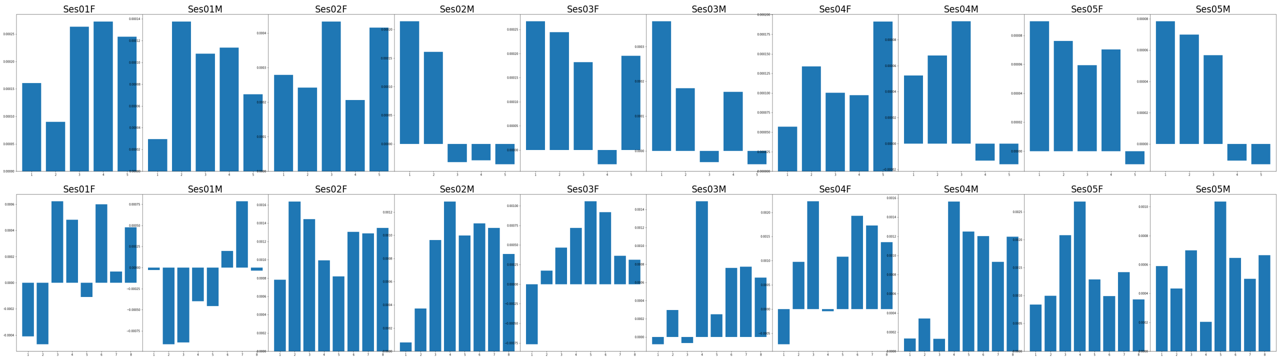

Fig. 2. *It shows the individual difference between accumulated attention values from PAaAN$_{PP}$ and A+T for five LLD categories and eight part of speech (POS) tags in the IEMOCAP database. The upper side distributions are from $\alpha_A$ and the lower ones are from $\alpha_T$.*

In terms of our proposed PAaAN architecture, where we integrate retrieved personality embedding derived from II-C2 in our recognition framework, we observe that the techniques of integration affects the recognition performances differently. As the results shown in Table I, *P-D* and *P-T* approaches obtain 67.5% and 62.4%, which are both inferior to our proposed *PAaAN$_{PP}$* with 3.70% and 12.18% relative degradation in accuracy. *P-D* and *P-T* approaches view the personal profile embedding as simply an auxiliary vector of information that is incorporated either at the level of final fully connected layer or within each modality-specific BLSTM time step. A simple concatenation in the deep neural network like *P-D* ignores the interaction between the personality embedding and the dynamically-varying audio and text behavior manifestation. On the other hand, concatenating the static personal embedding in each time step introduces unwanted redundancy, which deteriorates the recognition performances.

Lastly, we further investigate the necessity of acoustic domain adaptation in our experiments. We examine the performance obtained by ignoring domain adaptation, i.e., not using the learned SHLA mentioned in section II-C1. Specifically, we conduct the same cross corpus retrieval procedure in section II-C2 using the 1583 dimensional acoustic features without SHLA. The approach is denoted as *PAaAN$_{NA}$*. The resulting UAR drops to 67.9%, which declines a relative of 3.09% although it is still higher than $I_{A+T}$, the multimodal baseline without considering personal embedding. By computing the cosine similarity in the original feature space without adaptation, the identified retrievable personality embedding may be dramatically biased due to the natural cross corpora variability in the background noises, recording devices and scenario settings. In summary, we observe that indeed by integrating the target speaker's emotion profile through cross corpora retrievable personality embedding to the multimodal attention learning in the BLSTM architecture, it can obtain the state-of-the-art accuracy in the IEMOCAP database, i.e., 70.0% UAR in 4-class emotion classification; however, the domain adaptation technique is required to maintain its robustness in mitigating corpus-specific idiosyncratic factors.

### C. Analyses

In this section, we investigate the differences of accumulated attention values in our proposed PAaAN for both audio and text modality in the IEMOCAP database before and after integrating acoustic-based personality profile derived from the SSPNET. Since we have separate attention re-weighting vectors for audio and text, we elect to analyze this effect with respect to different major categories of frame level acoustic descriptors and part of speech (POS) tags for word level transcripts. The following is a list of acoustic LLDs and POS tags used:

- LLD (5 dimensions): loudness, MFCC, zero crossing rate, voice probability, F0
- POS (8 dimensions): coordinating conjunction (CC), preposition/subordinating conjunction (IN), adjective (JJ), modal (MD), noun (NN), personal pronoun (PRP), adverb (RB), verb (VB)

In this experiments, change of attention values is examined in the modality-wise salient parts of acoustic frames or lexical structures. In specifics, for the acoustic modality, we firstly compute the average values of each speaker on the five LLD categories as an threshold. Then, we accumulate the learned attention weights of the spoken frames with feature values that is over the threshold in each category as an indication of the salient acoustic frames. We can obtain accumulated attention weights before and after integrating personal embedding in our proposed *PAaAN$_{PP}$* on these salient frames. To analyze lexical modality, we accumulate the attention weights using similar procedure but directly based on the eight POS tags. Noted that each category of feature includes its associated sub-categories, e.g., loudness category contains the loudness LLD along with it's first and second derivatives sub-categories. Similarly, adjective comprise itself, comparative, and superlative types of adjectives. We use NLTK python toolbox to categorize words with POS tags [28].

The difference of accumulated attention weights between *PAaAN$_{PP}$* and $I_{A+T}$ in audio and text modalities are shown in Fig. 2. The figure depicts the ten speakers in the IEMOCAP database. It shows a varied distribution in both modalities across ten speakers. The first observation that we see is the distribution of changes is diverse across speakers reinforcing the known individual differences in the multimodal emotion expressions. Second, we see that generally, loudness and MFCC have increased attention weights while voice probability and F0 have decreased attention in 4 and 5 speakers respectively. Furthermore, the adverb category has an increased attention weights across all the speakers consistently. Most attention value decreases for category of coordinating conjunction, which is known to have less semantic meaning. The amount of increase and decrease, however, is subject-dependent.

Humans are capable of assessing the speech and language emotional messages conveyed by the other speakers once they become familiar with each other; for example, people with certain personality traits might tend to speak with higher pitch in their daily conversation. This personalized attention learning helps us reliably assess the emotional state of the others. In this work, we demonstrate through this analysis that our proposed *PAaAN* achieves such a personalized recognition with attention learning to re-estimate the important behavior regions for each target speaker individually by simultaneously considering the acoustic embeddings of personality with acoustic and lexical behaviors.

## IV. Conclusions

In this work, we propose to utilize a cross corpus acoustic-based personality retrieval approach within a PAaAN attention learning framework to integrate individual personality embedding for speech and language multimodal emotion recognition. Our proposed approach demonstrates a state-of-the-art 70% UAR in 4-class emotion classification task on the benchmark IEMOCAP database, and our further analysis on changes of attention before and after personality integration shows several important overall insights of individual personality's modulation on acoustic descriptors and part of speech tags though the effect largely varies from individual to individual. For the future work, we will immediately evaluate our PAaAN framework on other emotion corpora in order to evaluate the robustness of our framework. Technically, we will continue to investigate the effectiveness of other domain adaptation frameworks and different integration strategies through attention learning to ensure an efficient retrieval of relevant personal attributes from diverse corpora to move toward a personalized emotion recognition system.

## References

[1] N. Syam and A. Sharma, "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice," *Industrial Marketing Management*, vol. 69, pp. 135–146, 2018.

[2] M. B. Holbrook and R. Batra, "Assessing the Role of Emotions as Mediators of Consumer Responses to Advertising," *Journal of Consumer Research*, vol. 14, no. 3, pp. 404–420, 12 1987. [Online]. Available: https://doi.org/10.1086/209123

[3] B. Ferwerda, M. Schedl, and M. Tkalcic, "Personality & emotional states: Understanding users' music listening needs." CEUR-WS. org, 2015.

[4] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[5] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[6] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[7] S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, "Multimodal relational tensor network for sentiment and emotion classification," *arXiv preprint arXiv:1806.02923*, 2018.

[8] C. W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," *arXiv preprint arXiv:1805.06606*, 2018.

[9] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Interspeech 2018*, pp. 247–251, 2018.

[10] A. M. Kring and A. H. Gordon, "Sex differences in emotion: expression, experience, and physiology." *Journal of personality and social psychology*, vol. 74, no. 3, p. 686, 1998.

[11] S. T. Fiske, "Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain," *European Journal of Social Psychology*, vol. 30, no. 3, pp. 299–322, 2000.

[12] I. Ajzen, *Attitudes, personality, and behavior*. McGraw-Hill Education (UK), 2005.

[13] D. Matsumoto, "Are cultural differences in emotion regulation mediated by personality traits?" *Journal of Cross-Cultural Psychology*, vol. 37, no. 4, pp. 421–437, 2006.

[14] D. Byrne and L. Schulte, "Personality dispositions as mediators of sexual responses," *Annual review of sex research*, vol. 1, no. 1, pp. 93–117, 1990.

[15] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[16] W. Ng and E. Diener, "Personality differences in emotions: Does emotion regulation play a role?" *Journal of Individual Differences*, vol. 30, no. 2, pp. 100–106, 2009.

[17] N. Eisenberg, R. A. Fabes, M. Schaller, P. Miller, G. Carlo, R. Poulin, C. Shea, and R. Shell, "Personality and socialization correlates of vicarious emotional responding." *Journal of personality and social psychology*, vol. 61, no. 3, p. 459, 1991.

[18] L. Zhang, S. Peng, and S. Winkler, "Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship," *arXiv preprint arXiv:1811.08657*, 2018.

[19] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, p. 14, 2019.

[20] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 86–91.

[21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[22] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[25] C.-M. Chang and C.-C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5820–5824.

[26] C.-M. Chang, B.-H. Su, S.-C. Lin, J.-L. Li, and C.-C. Lee, “A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 377–382.

[27] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional mkl based multimodal emotion recognition and sentiment analysis,” in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[28] E. Loper and S. Bird, “Nltk: the natural language toolkit,” *arXiv preprint cs/0205028*, 2002.