Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions

Jinming Zhao
Renmin University of China
Haidian, Beijing, China
zhaojinming@ruc.edu.cn

Ruichen Li
Renmin University of China
Haidian, Beijing, China
ruichen@ruc.edu.cn

Shizhe Chen
Renmin University of China
Haidian, Beijing, China
cszhe1@ruc.edu.cn

Qin Jin*
Renmin University of China
Haidian, Beijing, China
qjin@ruc.edu.cn

ABSTRACT

Automatic emotion recognition is a challenging task which can make great impact on improving natural human computer interactions. In this paper, we present our solutions for the Cross-cultural Emotion Sub-challenge (CES) of Audio/Visual Emotion Challenge (AVEC) 2018. The videos were recorded in dyadic human-human interaction scenarios. In these complicated scenarios, a person's emotion state will be influenced by the interlocutor's behaviors, such as talking style/prosody, speech content, facial expression and body language. In this paper, we highlight two aspects of our solutions: 1) we explore multiple modalities's efficient deep learning features and use the LSTM network to capture the long-term temporal information. 2) we propose several multimodal interaction strategies to imitate the real interaction patterns for exploring which modality information of the interlocutor is effective, and we find the best interaction strategy which can make full use of the interlocutor's information. Our solutions achieve the best CCC performance of 0.704 and 0.783 on arousal and valence respectively on the challenge testing set of German, which significantly outperform the baseline system with corresponding CCC of 0.524 and 0.577 on arousal and valence, and which outperform the winner of the AVEC2017 with corresponding CCC of 0.675 and 0.756 on arousal and valence. The experimental results show that our proposed interaction strategies have strong generalization ability and can bring more robust performance.

KEYWORDS

Emotion Recognition, Domain Adaption, Multimodal, Dyadic Interaction

*Corresponding author.

AVEC'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266313

ACM Reference Format:
Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions. In *2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3266302.3266313

1 INTRODUCTION

Automatic emotion recognition is a crucial component to improve natural human-computer interactions. It has a wide range of applications in modern dyadic interaction scenarios, including call-center dialogue systems, conversational agents [12], depression severity prediction [11], mental health diagnoses and educational softwares [10].

Dimensional theories of emotion is one of the most popular computing models for emotion recognition [20]. It considers an emotion state as a point in a continuous space described by three dimensions corresponding to arousal (a measure of affective activation), valence (a measure of pleasure) and dominance (a measure of control). Therefore, dimensional emotion theory can model more subtle, complicated and continuous affective behaviors when compared to discrete theories of emotion.

The Cross-cultural Emotion Sub-challenge (CES) of Audio/Visual Emotion Challenge (AVEC) 2018 [11] is a dimensional emotion recognition task and provides a cross-cultural audio-visual dataset which is captured in the real-life dyadic human-human video chat scenarios. Previous works [3, 7–9] on dimensional emotion recognition tasks have explored different efficient multi-modal features, fusion methods and regression models. These features and models have proved quite effective, however fewer researches have focused on the influence from the interlocutor.

We [9] proposed a feature sequence construction combination strategy to imitate the interaction pattern and achieved significant improvement in AVEC2017, but we only considered the audio modality in previous work. In this work, we further propose several multimodal interaction strategies to make full use of the multimodal information of the interlocutor.

Our contributions to the challenge in this paper are from two aspects:

First, we investigate different efficient deep learning features from acoustic, visual and textual modalities, and we utilize Long Short-Term Memory (LSTM) [23] network to predict dimensional

continuous emotions. For the acoustic features, besides the expert-knowledge based features, we employ a more efficient audio representation model, VGGish [13], which is trained on a large scale dataset and can learn more richer audio representations. For the visual features, we compare the performance of two facial features extracted from different deep convolutional neural networks (CNN) [14, 24], which are pretrained on a facial expression recognition dataset, FER+ [2]. We also extract textual features from the pre-trained word embedding models [22]. Our experimental results show that the proposed deep learning features are effective and the fusion of different modality features can significantly improve the prediction performance on arousal, valence and likability. Previous works [9, 28] have proved that the LSTM network can capture long-term temporal information and achieve significant improvement than non-temporal models. So, in this paper, we use the LSTM network as our prediction model.

Second, we propose the real dyadic human-human interaction pattern under multimodal interaction scenarios. Based on this interaction pattern, we propose several efficient interaction strategies to take full advantage of the interlocutor's information. We explore and analyze the influence of the various modalities of the interlocutor through the proposed several interaction strategies, and find the best multimodal interaction strategy. Our proposed interaction strategies significantly improve the performance on arousal and valence on the challenge testing set of German language, which achieve CCC of 0.704 and 0.783 on arousal and valence respectively. Our experimental results indicate that the proposed interaction strategies have strong generalization ability and can achieve more robust performance.

The remainder of this paper is organized as follows: we present the related works and our proposed methods in Section 2 and Section 3, respectively. The experimental results and analyses are described in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED WORKS

Multimodal Features: In the past series of the AVEC challenge, some previous works have explored a variety of multi-modal features. Brady et al. [3], the winner of the AVEC2016 challenge, derived high-level acoustic, visual and physiological features from the low-level descriptors using sparse coding and deep learning methods. Chen et al. [9], the winner of the AVEC2017 challenge, focused on using deep CNNs to learn acoustic and facial expression features. The acoustic and facial expression features are extracted from the SoundNet [1] and the DenseNet [14] respectively, which bring significant performance improvement. Chen et al. [9] found the correlations between different emotion dimensions and fine-tune the DenseFace features extractor based on multitask learning method on the arousal and valence simultaneously. And the fine-tuned features gain significant improvement on both arousal and valence.

Multimodal Fusions: Emotion is naturally expressed through multiple modalities, such as acoustic, textual and visual modalities, which can capture complementary information about the emotions [21]. There are mainly three strategies to fuse the different modalities features, namely early-fusion, late fusion and model-level fusion [27]. Early fusion method, which concatenates different features as the input for the prediction model, is simple and can significantly improve performance [7, 9]. But the early fusion method suffers from the high dimensionality of the features and the synchronization of different features. Late fusion method combines the predictions from the different modalities by weighted sum or a second level model [15], however it ignores the interactions between different features. Model-level fusion method, a compromise for early and late fusion method, has been proposed to fuse the intermediate representations of the different features [6]. Recently, Chen et al. [8] proposed a temporal fusion model that can dynamically pay attention to relevant modality features through time, which shows improvements over the traditional fusion strategies.

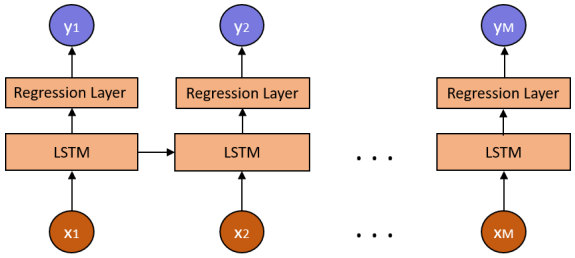

Figure 1: System Framework Overview. x refers to multi-modality features and y refers to emotion predictions.

Emotion Recognition Models: Previous works have proposed various context-sensitive models to capture the context information. Huang et al. [15] investigated the RML model with the annotation delay compensation and output associative fusion. Angeliki et al. [21] proposed a hierarchical HMM framework to fuse the context information and multimodal information for emotion prediction. Long Short Term Memory (LSTM) [23], one of the state-of-the-art sequence modeling techniques, has been widely and successfully applied in dimensional and continuous emotion recognition tasks [5, 7, 9].

Mutual Information in Dyadic Interaction: In dyadic human-human conversation scenarios, a person's emotion state will be influenced by the interlocutor's behaviors, such as talking style/prosody, speech content, facial expression and body language. Lee et al. [16] proposed that use Dynamic Bayesian Network (DBN) to explicitly model the conditional dependency between two interactive partner's emotions. Angeliki et al. [21] use a hierarchical framework which models emotional evolution. Soroosh et al. [19] presented a thorough analysis of IEMOCAP dataset [4] which reveals that 72% conversational partners present similar emotions. Based on this finding, they proposed a novel cross-modality, cross-speaker emotion recognition method to improve the recognition performance. Chen et al. [9] proposed an interaction strategy that using feature sequence construction combinations to imitate the interaction patterns on the audio modality for continuous emotion prediction, which achieves significant performance improvement.

3 PROPOSED METHODS

3.1 System Framework

In this paper, we adopt the LSTM [23] network as our prediction model for capturing the long-term temporal information. As is

shown in Fig. 1, given an feature sequence $x = \{x_1, x_2, ..., x_M\}$ as the system input, where M is the max time step of the LSTM model. A regression layer follows the output of the LSTM layer, and then we get the continuous emotion predictions, $y = \{y_1, y_2, ..., y_M\}$.

We use the mean square error (MSE) as our loss function, which minimizes:

$$L = \frac{1}{2T}\sum_{t=1}^{T}(g_t - y_t)^2 \quad (1)$$

where g_t is the ground truth emotion label of the t time step, and T is the total time steps of the data.

3.2 Multimodal Features

VGGish We extract short-term acoustic features from the VGGish [13] model which is trained on a large scale dataset and can learn more richer audio representations. The recordings are first divided into non-overlapping frames with window size 0.98s. Each frame is then transformed into log-mel spectrogram features as the input of the VGGish model. We extract activations from the last fully connected layer with dimensionality of 128 as the audio embedding features and refer the features as "vggish.100ms".

Word Vectors Word vectors are distributional word representations learned from massive textual dataset [22], which are not only more compact than the Bag of Words (BOW) representations but also are related to the semantic meanings of the words. We adopt an unofficial pretrained German word embedding model [1] with 300 embedding dimensions and mean pool the word embeddings in the turn as turn-level features. Since the quality of the German word embedding model might not be very robust, we also translate the German transcriptions into English by Google Translator and use the Google English word embedding model [2] to extract textual features. Additionally, for Hungarian transcriptions, we first translate it to German and English, ant then extract textual features from the above mentioned German and English word embedding models.

VGG-style CNN (VGGFace): The VGGFace model has the same structure used in [9] and pretrained on the FER+ dataset [2]. Though the target is different between the VGGFace model and the AVEC affective task, the middle layers can contain useful features related to the facial expressions. So we extract facial expression features from the conv5 and conv6 layers as the frame-level features, and refer the features as "vggface.conv5" and "vggface.conv6" respectively.

DenseNet-style CNN (DenseFace): The recent proposed Densely Connected Convolutional Networks (DenseNet) [14] have achieved the state-of-the-art performance in many image recognition tasks. It connects all the preceding layers as the input for a certain layer, which can strengthen feature propagation and alleviate the gradient vanishing problem. Besides, due to the feature reuse, DenseNet only needs to learn a small set of new feature maps in each layer and thus requires fewer parameters than traditional CNNs, which is more suitable for small datasets. The details of the DenseNet structure, training process and the network finetuning are described in [9]. We extract the activations from the last mean pooling layer of the finetuned model and refer the feature as "denseface.tune". Further, we finetune the DenseNet model based on the multi-task learning method on arousal and valence simultaneously which is the same as reported in [9], and the features extracted from the finetuned DenseNet model are referred as "denseface.tune.AV". To match with the shift of ground-truth labels, we apply mean pooling over consecutive 5 CNN frame features. The frames where no face is detected are filled with zeros.

[1] http://devmount.github.io/GermanWordEmbeddings
[2] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

3.3 Proposed Interaction Strategies

Motivated by the discovery that of the influence of the similarity of the emotions between the partners [19]. We first compute the correlations of continuous targets of each pair in the AVEC2018 dataset and the distribution of the correlation of the pairs is shown in Fig. 3. There are 54.17% pairs and 70.83% pairs have strong positive correlation on arousal and valence respectively, which indicates strong mutual influence in their expressive behaviors.

In the dyadic human-human conversation scenarios, when one person talks, the another is in silence, alternately. When one person talks, he/she has audio, text and facial expression information. The interlocutor, meanwhile, is silent, he/she has facial expression information only. Accordingly, we proposed the real dyadic human-human interaction pattern shown in in Fig. 2. As shown in Fig. 2, for speaker's continuous emotion evolution, the influences are mainly derived from three aspects: the previous speaker's context information, the previous interlocutor's context information and the current interlocutor's information. Motivated by the interaction strategy of the acoustic modality mentioned in [9], we propose several unimodal and multimodal interaction strategies to imitate the multimodal interaction pattern.

In our notation, we use S_t^a, S_t^f and S_t^t to represent the acoustic, facial expression, textual modality features of the speaker at t time step. We use the SZ_t^a to represent speaker's acoustic features which are filled with zeros at t time step. As for interlocutor's features, we just replace the S in above defined symbols with I, for example, I_t^a represent the audio modality feature of the interlocutor at t time step.

Audio Interaction We explore the influence of the interlocutor's behaviors on speaker's emotions with the audio modality only. As is shown in Fig. 2, when in $T-1$ turn, the speaker is talking and the interlocutor is silent. Therefore, in this turn, the speaker only has audio information, and the interlocutor does not have any information. When in T turn, the speaker is silent and the interlocutor is talking. In this turn, the speaker does not have any information, and the interlocutor has audio information only. So, the combination of feature sequence, which is similar to the "Double" feature sequence construction combination mentioned in [9], at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; IZ_t^a] & if\ t \in T-1 \\ [SZ_t^a; I_t^a] & if\ t \in T \end{cases} \quad (2)$$

Facial Expression Interaction We explore the influence of the interlocutor's behaviors on speaker's emotions with the visual modality only. Unlike the consideration of the existence of acoustic modality only, whether in $T-1$ turn or in T turn, the speaker and interlocutor both have facial expression information. Note that we ignore the interlocutor's facial expression information when the speaker is silent. There are two reasons for this design: 1. when speaker talks, the interlocutor's facial expression information can make up for the missing audio information. However, when

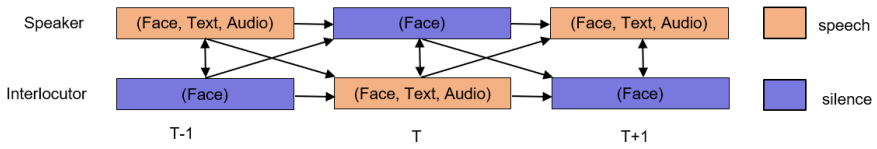

Figure 2: Interaction Framework Overview. Each block denotes a turn which is defined as the portion where speech belonging to a single speaker before he/she finishes speaking, and may consists of multiple original segmented utterances. The orange block denotes the speech turn and purple block denotes the silence turn. "T" represents the turn time step.

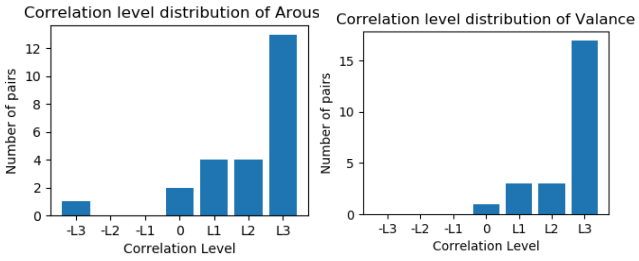

Figure 3: Correlation Distribution. The map from correlation levels to PCC is: {L3:[0.5, 1.0], L2:[0.3, 0.5], L1:[0.1, 0.3],0:[-0.1, 0.1], -L1:[-0.3, 0.1], -L2:[-0.5,-0.3], -L3:[-1.0,-0.5]}.

speaker is silent, the speaker just has facial expression information and the targets were labeled only by speaker's facial expression. 2. furthermore, the interlocutor's facial expression may not be exactly accurate when he/she talks. So, the feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^f; I_t^f] & if\ t \in T-1 \\ [S_t^f; IZ_t^f] & if\ t \in T \end{cases} \quad (3)$$

Multimodal Baseline We first present our multimodal baseline system which only consider the audio and facial expression information of the speaker. As is shown in Fig. 2, when in $T-1$ turn, the speaker is talking and has both audio and facial expression information. when in T turn, the speaker is silent and has only facial expression information. The feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^f] & if\ t \in T-1 \\ [SZ_t^a; S_t^f] & if\ t \in T \end{cases} \quad (4)$$

AFA Interaction Based on the "Multimodal Baseline" strategy, we consider the the influence of the interlocutor's audio information on the speaker's emotions. As is shown in Fig. 2, the interlocutor is silent in $T-1$ and is talking in T turn. So, the feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^f; IZ_t^a] & if\ t \in T-1 \\ [SZ_t^a; S_t^f; I_t^a] & if\ t \in T \end{cases} \quad (5)$$

AFF Interaction Then, based on the "Multimodal Baseline" strategy, we consider the the influence of the interlocutor's facial expression information on the speaker's emotions. Since the reasons mentioned in "Facial Expression Interaction", we ignore the interlocutor's facial expression information when the speaker is silent. So, the feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^f; I_t^f] & if\ t \in T-1 \\ [SZ_t^a; S_t^f; IZ_t^f] & if\ t \in T \end{cases} \quad (6)$$

AFAF Interaction Then, based on the interaction strategies mentioned above, we utilize both audio and facial expression information of the interlocutor. The feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^f; IZ_t^a; I_t^f] & if\ t \in T-1 \\ [SZ_t^a; S_t^f; I_t^a; IZ_t^f] & if\ t \in T \end{cases} \quad (7)$$

ATFATF Interaction Finally, we utilize all available modalities, including acoustic, textural and visual modalities. Since the text is translated from the audio, this interaction strategy is similar to the "AFAF Interaction" strategy. the feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^t; S_t^f; IZ_t^a; IZ_t^t; I_t^f] & if\ t \in T-1 \\ [SZ_t^a; SZ_t^t; S_t^f; I_t^a, I_t^t; IZ_t^f] & if\ t \in T \end{cases} \quad (8)$$

ATFAT Interaction Additionally, based on the experimental results of the "Multimodal Baseline" and the "AFF Interaction" strategies shown in Table 4, the facial expression information of the interlocutor slightly drop the performance. So, based on the "ATFATF Interaction" strategy, we ignore the facial expression information of the interlocutor. And the feature sequence combination at t time step can be denoted as F_t:

$$Ft = \begin{cases} [S_t^a; S_t^t; S_t^f; IZ_t^a; IZ_t^t] & if\ t \in T-1 \\ [SZ_t^a; SZ_t^t; S_t^f; I_t^a, I_t^t] & if\ t \in T \end{cases} \quad (9)$$

4 EXPERIMENTS

4.1 Corpus Description

In this paper, we use the AVEC 2018 corpus which is a subset of the Sentiment Analysis in the Wild (SEWA) dataset [3]. There are 64 German subjects and 66 Hungarian subjects in the corpus. Subjects participated in pairs and were asked to discuss the commercial products through the video chat. All audio-visual recordings were collected "in-the-wild" using standard webcams and microphones in the subjects offices or homes. The duration of each conversation is about 3 minutes. All three emotion dimensions are annotated every 100ms and scaled into [-1, +1]. The detailed data distribution of this corpus is shown in Table 1.

[3] http://sewaproject.eu

Table 1: The corpus distribution. "Train_DE", "Val_DE", "Test_DE" and "Test_HU" denote the training set of German language, the validation set of German language, the testing set of German language and the testing set of Hungarian language respectively.

Set	Subjects	Culture
Train_DE	34	German
Val_DE	14	German
Test_DE	16	German
Test_HU	66	Hungarian

4.2 Experimental Setup

We implement the LSTM network proposed in [23] with the tensorflow[4] deep learning toolkit. The number of layers in the LSTM is set to be 1 and the hidden units number is optimized for different input features. We use the truncated back propagation through time (BPTT) with max time step of 100 to train our LSTM networks. We train at most 120 epochs for each model. Dropout is adopted to avoid overfitting with dropout rate of 0.5. Adam optimizer is applied and the learning rate is initialized from 0.01 and reduce half every 50 epochs. The predictions of our models are smoothed by simply averaging the predictions within a fixed window. The concordance correlation coefficient (CCC) [17] works as the evaluation metric for this challenge, which is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{10}$$

where μ_x and μ_y are the means of the sequence x and y, and σ_x and σ_y are the corresponding standard deviations. ρ is the Pearson Correlation Coefficient (PCC) between x and y.

Table 2: CCC performance of different unimodal features on the validation set.

	Dim	A	V	L
baseline-audio	–	0.421	0.398	0.15
vggish.100ms	120	0.4791	0.4473	0.1038
vggish.100ms.empty	120	**0.6041**	**0.5107**	**0.1559**
baseline-video	–	0.486	0.549	**0.212**
vggface.conv5	120	0.6224	**0.7006**	0.1658
vggface.conv6	120	0.5327	0.5937	0.0791
denseface	120	0.5138	0.6753	0.1136
denseface.tune.AV	120	**0.6897**	0.6913	0.2107
english-wordvec	120	0.4691	0.517	0.3942
german-wordvec	120	**0.4866**	**0.5603**	**0.4356**

4.3 Unimodal Results

We use the LSTM network as our prediction model and test the effectiveness of the proposed deep learning features. As is shown in Table 2, the performance of the proposed "vggish.100ms" audio features outperform the official baseline audio features [11] on arousal and valence. The official audio recordings contain both speaker's and interlocutor's audio signals, therefore directly use the original audio recordings is not accurate. By comparing the performance of "vggish.100ms" and "vggish.100ms.empty", where "empty" in "vggish.100ms.empty" means removing the interlocutor's speech segment by the official turn information, the "vggish.100ms.empty" features significantly outperform the "vggish.100ms" performance. The experimental results show that the official combination method, which simply concatenates the speaker's audio signals and the interlocutor's audio signals in time dimension, not only misses the advantage of the interlocutor's information, but also brings a lot of noises that result in worse performance.

For the performance of the visual short-time features, the "vggface.conv5" features generalize better from the categorical to dimensional emotion recognition than the "vggface.conv6" features. Due to the structure of the DenseNet model, the DenseFace features do not perform as well as the VGGFace features. Motivated by [9], we use multi-task learning method to finetune the DenseNet model with arousal and valence targets together, and the finetuned features achieve significant performance improvement.

For the performance of the text modality, due of the data insufficiency and the errors occurred in the recognition and transcription preprocess stages, the textual features get the worst performance on arousal and valence. However, the textural features achieve the best performance on likability, while audio and visual features perform extremely bad on likability. We suggest the reason of this phenomenon is that the likability might be mainly reflected from speech contents rather than the audio-visual signals.

4.4 Multimodal Results

We use the early fusion strategy to explore the complementarity of the multimodal and different features. As is shown in Table 3, the results of multimodal fusion features all improve the performance of the best unimodal features except on likability. Based on the "denseface.tune.AV" features which perform well on both arousal and valence, when fused with the textual features together, and the fused features can bring performance gains on all three dimensions. From the fusion of the "denseface.tune.AV" features and the "vggish.100ms.empty" features, we get significant improvement on arousal and valence compared to the fusion of "denseface.tune.AV" and "german-wordvec", but get poor result on likability. These results show that the text information is the key factor for the prediction of likability, and the audio information can bring complementary information for the prediction of arousal and valence. Further, compared with the "denseface.tune.AV-vggish.100ms.empty" features, we fuse the "denseface.tune.AV", "vggish.100ms.empty" and "german-wordvec" together which result in comparable performance on arousal, some performance improvement on valence, and a bit performance drop on likability which indicate that the audio information will bring noises for the prediction of likability. Finally, the "vggish.100ms.empty-denseface.tune.AV-vggface.conv5-german-wordvec-english-wordvec" features, the fusion of five different features from three modalities, achieve the best multimodal result on all arousal, valence and likability three dimensions.

[4] https://www.tensorflow.org

Table 3: CCC performance of different multimodal fusion features on the validation set.

	Dim	A	V	L
unimodal best results	120	0.6897	0.7006	0.4356
denseface.tune.AV-german-wordvec	120	0.7123	0.7389	0.4717
denseface.tune.AV-vggish.100ms.empty	240	0.7449	0.7534	0.1468
denseface.tune.AV-vggish.100ms.empty-german-wordvec	240	0.7444	0.7618	0.354
vggish.100ms.empty-denseface.tune.AV-vggface.conv5	240	0.7692	0.7599	0.2869
vggish.100ms.empty-denseface.tune.AV-vggface.conv5-german-wordvec-english-wordvec	480	**0.7914**	**0.7823**	**0.5098**

Table 4: CCC performance of different interaction strategies on the validation set. We use this "feature name (interaction strategy)" format for uniform naming, for example, "vggish.100ms.interact (Audio Interaction)" denotes "vggish.100ms.interact" features based on the "Audio Interaction" strategy defined in Section 3.3.

	Dim	A	V	L
vggish.100ms.empty	120	0.6041	0.5107	0.1559
vggish.100ms.interact (Audio Interaction)	120	**0.6495**	**0.5241**	**0.2946**
denseface.tune.AV	120	0.6897	**0.6913**	0.2107
denseface.tune.AV.interact (Facial Expression Interaction)	240	**0.7339**	0.6868	**0.3371**
denseface.tune.AV-vggish.100ms.empty (Multimodal Baseline)	120	0.7449	0.7534	0.1468
denseface.tune.AV-vggish.100ms.AFF (AFF Interaction)	240	0.7413	0.7307	0.2563
denseface.tune.AV-vggish.100ms.AFA (AFA Interaction)	240	**0.7764**	**0.7645**	0.2809
denseface.tune.AV-vggish.100ms.AFAF (AFAF Interaction)	480	0.7646	0.7392	**0.4608**
denseface.tune.AV-vggish.100ms.empty-german-wordvec	240	0.7444	**0.7618**	0.354
denseface.tune.AV-vggish.100ms-german-wordvec.ATFAT (ATFAT Interaction)	600	0.7433	0.7589	0.3707
denseface.tune.AV-vggish.100ms-german-wordvec.ATFATF (ATFATF Interaction)	600	**0.7882**	0.711	**0.3866**

Table 5: CCC performance of multimodal interaction strategies on the validation set and testing sets. "multimodal-based" systems denotes the fusion features which without using interlocutor's information and "multimodal-interact-based" systems denote the fusion features are processed with our proposed interaction strategies.

		Val-DE	Test-DE	Test-HU
A	multimodal-based	**0.820**	0.662	0.562
	multimodal-interact-based	0.789	**0.704**	0.540
V	multimodal-based	**0.795**	0.755	0.438
	multimodal-interact-based	0.770	**0.783**	0.421
L	text-based	0.496	0.433	0.194
	text-interact-based	0.447	0.405	0.193

4.5 Interaction Results

First, we use two our proposed unimodal interaction strategies to verify the interlocutor's influence under unimodal conditions. In Table 4, the unimodal results show that using the interlocutor's information can significantly improve the performance, except that the "denseface.tune.AV.interact" features have comparable performance with the "denseface.tune.AV" on valence. These results suggest that the interlocutor's audio and facial expression information are helpful for predicting the speaker's emotions under unimodal conditions.

Then, under the multimodal conditions, we use our proposed multimodal interaction strategies to verify the interlocutor's influence and to explore which modalities information of the interlocutor are helpful for the prediction of speaker's emotions. Compared with the performance of the "denseface.tune.AV-vggish.100ms.empty" features shown in Table 4, the "denseface.tune.AV-vggish.100ms.AFF" features result in relative poor performance on arousal and valence, which indicates that the facial expression information of the interlocutor can not provide useful information for the prediction of speaker's emotion under multimodal conditions. Furthermore, by comparing the results of "denseface.tune.AV-vggish.100ms.AFAF" features and "denseface.tune.AV-vggish.100ms.AFA" features, we can get the same conclusion. The reason of this phenomenon is that interlocutor's facial expression is relatively less obvious and may inhibit the expression of emotions of the speaker. The "AFAF Interaction" and "AFA Interaction" strategies both significantly outperform the "Multimodal Baseline" strategy, and the "AFA Interaction" strategy achieve the best result on arousal and valence under the condition of considering only audio and facial expression two modalities. These results show that the audio information of the interlocutor can bring positive impact for the prediction of the speaker's emotions on both arousal and valence.

The "Interact.ATFATF" strategy which utilizes all available modalities results in best result on arousal on the validation set. However, the result of "Interact.ATFATF" strategy is lower than the others multimodal features with or without adopting the interaction strategies on valence. And the "Interact.ATFAT" strategy, which ignores the facial expression information of the interlocutor,

did not result in the expected performances which are better than the performance of the others experiments which considering three modalities. We guess that the main reason of this phenomenon is that the text information is not rich and not accurate. Another possible reason of this phenomenon is that the high dimensionality of the interaction construction features and the LSTM network can't handle this complex interaction strategies. In the future, we will explore more efficient interaction models to handle more complex scenarios.

As we can see from the Table 4, all the interaction strategies outperform the others which do not consider the interlocutor's influence on likability.

Finally, we verify the efficiency of our proposed interaction strategies on the challenge testing sets. As is shown in Table 5, the multimodal based features outperform the multimodal interaction features on the validation set. However, the multimodal interaction features significantly outperform the multimodal based features on both arousal and valence and result in the best result on the arousal and valence on the testing set of German culture. These results indicate that our proposed interaction strategies have stronger generalization ability and can bring more robust performance. For the prediction on likability, we find that only the textual features perform better than the multimodal features, though the multimodal have far better performance on the validation set, which is similar to the observation as reported in the winner of the AVEC2017 challenge conclusion mentioned in Chen et al. [9]. Our proposed methods do not show the expected results on the likability, a possible reason is that the text information is not rich enough and may contain errors.

Table 6: CCC performance of best results of five submissions on the validation set and testing sets.

	A	V	L
Val-DE	0.820	0.795	0.549
Baseline-Dev-DE [11]	0.581	0.649	0.288
Test-DE	0.704	0.783	0.433
Baseline-Test-DE [11]	0.524	0.577	0.038
Test-HU	0.562	0.438	0.197
Baseline-Test-HU [11]	0.436	0.405	0.023

4.6 Submission Results

For the prediction on arousal and valence, we average the prediction of multiple multimodal systems. However, for the prediction on likability, the best performance is achieved with the textual features only. We also scale and shift the predictions according to the statistics on the validation set.

As is shown in Table 6, our proposed interaction strategies achieve the best result on arousal and valence for German cultural. This result significantly outperforms the official baseline performance. Furthermore, our solutions achieve the CCC of 0.704, 0.783 on arousal and valence respectively on the challenge testing set of German, which outperforms the winner of the AVEC2017 with corresponding CCC of 0.675 and 0.756 on arousal and valence.

There is big gap of performance between the testing set of German culture and Hungarian culture, we have explored several unsupervised domain adaption methods [18, 25, 26], due to the time constraints, these methods have not worked well. In the future, we will continue to explore some unsupervised domain adaption methods for the cross-culture challenge.

5 CONCLUSIONS

In this paper, we explored different efficient deep learning features of all available modalities including the acoustic, visual and textual modalities, and we proposed several interaction strategies under unimodal scenarios and mutilmodal scenarios, and achieve significant performance improvement. The audio and facial features are extracted from the pretrained VGGish and DenseNet models and we use the LSTM network which is used to capture the long-term temporal information, as our prediction model. According to the real dyadic human-human interactions patterns, we proposed several efficient interaction strategies for both unimodal and multimodal scenarios, and bring performance gains under these scenarios. The experimental results suggest that multimodal interaction features significantly outperform the multimodal features which do not consider the interlocutor's influence on both arousal and valence on testing set of German culture. The results also indicate that our proposed interaction strategies have stronger generalization ability and can bring more robust performance. In the future, we will explore more efficient features for the prediction of likability, more efficient interaction model for the multimodal dyadic conversation scenarios and unsupervised domain adaption methods for the cross-cultural dataset.

ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202. This work is also partially supported by National Natural Science Foundation of China (Grant No. 61772535).

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. (2016).
[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction*. 279–283.
[3] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S. Huang. 2016. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In *International Workshop on Audio/visual Emotion Challenge*. 97–104.
[4] Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources Evaluation* 42, 4 (2008), 335–359.
[5] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2015. Long Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC '15)*. ACM, New York, NY, USA, 65–72. https://doi.org/10.1145/2808196.2811634
[6] Jun Kai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2014. Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning. In *International Conference on Multimodal Interaction*. 508–513.
[7] Shizhe Chen and Qin Jin. 2015. Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks. In *International Workshop on Audio/visual Emotion Challenge*. 49–56.

[8] Shizhe Chen and Qin Jin. 2016. Multi-modal Conditional Attention Fusion for Dimensional Emotion Prediction. In *ACM on Multimedia Conference*. 571–575.

[9] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *The Workshop on Audio/visual Emotion Challenge*. 19–26.

[10] Cristina Conati. 2002. Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence* 16, 7-8 (2002), 555–575.

[11] Fabien Ringeval and Björn Schuller and Michel Valstar and Roddy Cowie and Heysem Kaya and Maximilian Schmitt and Shahin Amiriparian and Nicholas Cummins and Denis Lalanne and Adrien Michaud and Elvan Çiftçi and Hüseyin Güleç and Albert Ali Salah and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*, Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic (Eds.). ACM, Seoul, Korea.

[12] N Fragopanagos and J. G. Taylor. 2002. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (2002), 32–80.

[13] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, and Bryan Seybold. 2016. CNN architectures for large-scale audio classification. (2016), 131–135.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2261–2269.

[15] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. 2015. An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. In *International Workshop on AVECe*. 41–48.

[16] Chi Chun Lee, Carlos Busso, Sungbok Lee, and et.al. 2009. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *INTERSPEECH*. 1983–1986.

[17] L. I. Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989).

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. (2015), 97–105.

[19] Soroosh Mariooryad and Carlos Busso. 2013. Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition. *IEEE Transactions on Affective Computing* (2013).

[20] Stacy Marsella and Jonathan Gratch. 2014. Computationally modeling human emotion. *Communications of the Acm* 57, 12 (2014), 56–67.

[21] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. 2012. A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs. In *ICASSP*. 2401–2404. https://doi.org/10.1109/ICASSP.2012.6288399

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 26 (2013), 3111–3119.

[23] Ha?im Sak, Andrew Senior, and Fran?oise Beaufays. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *Computer Science* (2014), 338–342.

[24] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

[25] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial Discriminative Domain Adaptation. (2017).

[26] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. *Computer Science* (2014).

[27] Chung Hsien Wu, Jen Chun Lin, and Wen Li Wei. 2014. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *Apsipa Transactions on Signal Information Processing* 3 (2014), –.

[28] Martin WÃűllmer, Angeliki Metallinou, Florian Eyben, BjÃűrn Schuller, and Shrikanth S. Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September*. 2362–2365.