Interspeech 2018
2-6 September 2018, Hyderabad

Encoding Individual Acoustic Features using Dyad-Augmented Deep Variational Representations for Dialog-level Emotion Recognition

Jeng-Lin Li[1,2] *and Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

cllee@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Face-to-face dyadic spoken dialog is a fundamental unit of human interaction. Despite numerous empirical evidences in demonstrating interlocutor's behavior dependency in dyadic interactions, few technical works exist in leveraging the unique pattern of dynamics in task of advancing emotion recognition during face-to-face settings. In this work, we propose a framework of encoding an individual's acoustic features with dyad-augmented deep networks. The dyad-augmented deep networks includes a general variational deep Gaussian Mixture embedding network and a dyad-specific fine-tuned network. Our framework utilizes the augmented dyad-specific feature space to incorporate the unique behavior pattern emerged when two people interact. We perform dialog-level emotion regression tasks in both the CreativeIT and the NNIME databases. We obtain affect regression accuracy of 0.544 and 0.387 for activation and valence in the CreativeIT database (a relative improvement of 4.41% and 4.03% compared to using features without augmenting the dyad-specific representation), and we obtain 0.700 and 0.604 (4.48% and 4.14% relative improvement) for regressing activation and valence in the NNIME database.

Index Terms: variational deep embedding, dyadic interaction, emotion recognition, feature augmentation, frozen fine-tuning

1. Introduction

Dyadic interaction is a basic unit of face-to-face human communication providing an important gateway for humans to convey emotion, exchange information, and foster mutual understanding [1]. Interpersonal dependency between interlocutors emerges naturally during interactions. This dependency is evident both in their internal states (e.g., emotion, cognition, perception, etc) and their expressive behaviors (e.g., speech, language, visual expressions, etc). The mutual behavioral dependencies during interactions have been well-studied in the field of psychology, i.e., known varying as synchrony, entrainment or adaptation [2, 3]. Preliminary research have also formulated this inter-dependency as a system framework to quantitatively interpret the nature of interactions [4, 5]. This naturally-occurring dependencies internally and behaviorally between the interacting interlocutors over time lead to unique intricate dyad-specific patterns of interaction dynamics [6].

As the next-generation human-centered applications become more prevalent, robust and reliable affective sensing in *face-to-face* interactions is becoming more critical, especially important in supporting technologies of natural dialog interfaces, human behavior understanding [7], and health applications [8]. While tremendous effort has been developed in emotion recognition, majority of these research has focused mainly on developing algorithms for an individual's behavior in isolation (e.g., [9, 10, 11]). Only recently, some research have started to leverage the inter-dependencies between interlocutors to improve affect recognition of an individual. For example, Yang et al. conducted research in computational modeling of analyzing joint behavior dynamics between dyads as a function of their emotional states [12, 13, 14]; Metallinou proposed a hierarchical emotion evolution model [15, 16], and also several other similar research [17, 18] together have demonstrated the usefulness of integrating dyadic patterns of affective states in task of individual's affect recognition. Most of these recent works have presented frameworks that explicitly model the subtle and intricate inter-dependency at the level of emotional states not directly at the level of behavior representations.

In this work, we propose a novel network architecture to obtain robust acoustic representation for an individual during dyadic interactions. The approach includes two major components: a general representation and a dyad-specific dynamic representation. In specifics, our framework consists of using variational autoencoder jointly learned with mixture of Gaussian prior at the latent layer, i.e., variational deep embedding (VaDE) [19]; this encoder network can be used to derive general acoustic behavior representation. The modeling of intricate behavior patterns emerges during dyadic face-to-face interactions can be cast as learning dyad-specific network by adapting the general model to the specific dyad. Due to the subtlety of these behavior dynamics, we utilize a encoder-decoder *frozen* adaptation strategy that only updates the middle generative latent layer to mitigate the issue of forgetting effect. With these two dyad-augmented deep variational autoencoder networks, we can represent an individual acoustic features as a general VaDE representation augmented with the dyad-specific representation.

We evaluate our proposed framework for the task of dialog-level emotion recognition in two different databases: the CreativeIT (CIT) [20] and the NNIME database [21]. In specifics, we obtain dialog-level affect regression accuracy of 0.544 and 0.387 for activation and valence in the CIT database (a relative improvement of 4.41% and 4.03% compared encoding network without augmenting dyad-specific representation). We obtain 0.700 and 0.604 (4.48% and 4.14% relative improvement) for regressing dialog-level activation and valence in the NNIME database. The rest of the paper is organized as follows: research framework is in Section 2 followed by experiment setup and results (Section 3); conclusion is presented in Section 4.

2. Research Methodology

2.1. Emotion Databases

We utilize two dyadic emotion interaction databases, the NNIME and the CIT, in this work. We will briefly describe each database in the following section.

10.21437/Interspeech.2018-1455

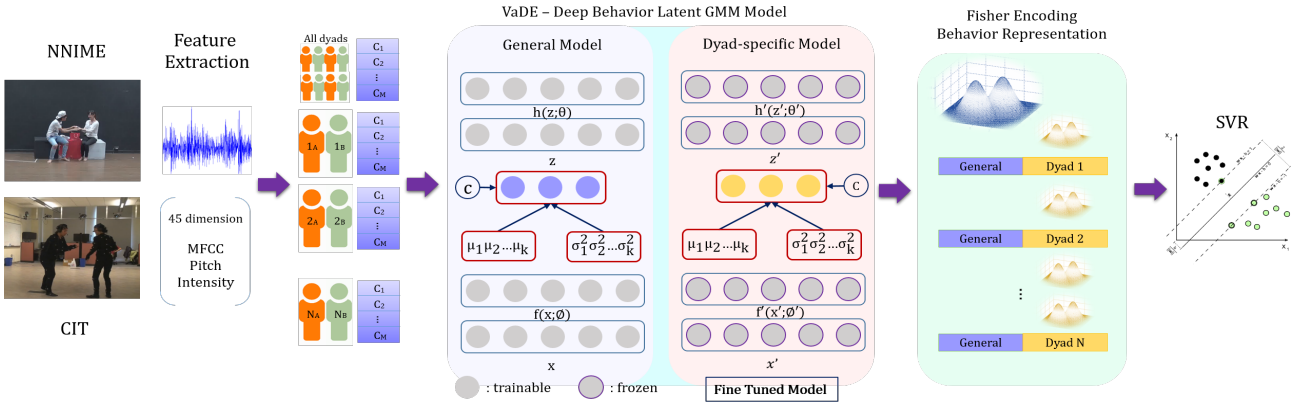

Figure 1: *This is the overall framework for an individual's dialog-level emotion recognition. We first extract low-level descriptors. Then, the LLDs are encoded using two networks of general VaDE and dyad-specific VaDE. General representation acts as a behavior representation learned from the entire database while dyad-specific representation embeds dyadic interaction dynamics. Our proposed dyad-augmented representation is a concatenation of these two representations after performing dialog-level Fisher encoding. The general VaDE is composed by encoding network f and decoding network h with parameters ϕ and θ. Similarly, the dyad-specific VaDE has encoding network f' and decoding network h' with parameters ϕ' and θ' respectively.*

2.1.1. *The USC CreativeIT Database (CIT)*

The CreativeIT database (CIT) is a publicly-available multimodal emotional corpus consisting of dyadic affective interactions [20]. The database has previously been used in studies of dyadic affective behavior interplay [12, 13, 14]. The dyadic interactions are carried out using an established theatrical acting technique, i.e., Active Analysis, to elicit natural affective behaviors. There are 7 unique male-female pairs (14 speakers) of individuals engage in approximately 3-minute long affective interactions (a total of 40 sessions, 80 total annotation samples). The audio recordings of each individual are collected from lapel microphones. Dialog-level emotion labels of each subject for every interactions are rated using dimensional attributes, i.e., activation and valence, on a scale of [1, 5] by at least 3 naive raters. In this work, the average ratings serve as ground truths.

2.1.2. *The NNIME Emotion Database (NNIME)*

The NNIME emotion database is a recently-published Chinese dyadic multimodal interaction corpus using a similar setup as the USC CreativeIT database [21]. The affective dyadic interactions are hypothesized to be in real life scenarios with an overall targeted affective atmosphere. The naturalness in the affective behavior display is further ensured by a professional director. There are 99 interaction sessions (198 annotation samples) with each lasts around 3 minutes long. There are a total of 22 unique dyads (44 individuals) in the database. The interaction sessions are recorded by a video camera facing the stage and lapel microphone placing on each individual. The emotion attributes (dialog-level) of activation and valence on a scale of [1, 5] are annotated by 42 raters regarding the perceived emotion from each individual. In our work, we take the average of the 42 ratings as our ground truths.

2.2. Dyad-Augmented Deep Variational Representations

We propose to encode acoustic features using dyad-augmented deep variational representations. The overall computational framework, including acoustic low-level descriptors, VaDE representation, and dyad-augmented representation, is depicted in Fig. 1. Each will be described in detail further below.

2.2.1. *Acoustic Low-level Descriptors*

We extract 45 acoustic low-level descriptors (LLDs) at the frame level (25ms window 10ms step-size) over the speaking region. The LLDs include 13 Mel Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and their associated delta and delta-delta. All of the extracted LLDs are further z-normalized with respect to each individual speaker.

2.2.2. *Variational Deep Embedding Network (VaDE)*

In our framework, the representations are learned from variational deep embedding (VaDE) model [19], which is a variational autoencoder(VAE) with Gaussian mixture prior. The use of Gaussian mixture relaxes the assumption and alleviate over-regularization problem of a standard VAE with a single Gaussian prior distribution [22]. VaDE is learned by maximizing the evidence lower bound (ELBO), $\mathcal{L}_{EMBO}$, of log-likelihood function, $p(x)$, for the input sample x and latent factor z:

$$\begin{aligned} \log p(x) &= \log \int_z \sum_c p(x,z,c)dz \\ &\geq E_{q(z,c|x)}\left[\log \frac{p(x,z,c)}{q(z,c|x)}\right] = \mathcal{L}_{EMBO} \end{aligned} \tag{1}$$

where c indicates hidden clustered (mixture) states $\in \{1,\ldots,K\}$ and the $q(z,c|x)$ is the variational posterior approximation distribution of the true posterior $p(z,c|x)$, which can be factorized as: $q(z,c|x) = q(z|x)q(c|x)$. To model $q(z|x)$, the encoding network $f(x;\phi)$ is used to jointly learn the latent GMM parameters λ (weight, mean and variance matrix denoted as π_c, μ_c and σ_c^2 for the c-th cluster.

$$[\tilde{\mu}, \log \tilde{\sigma}^2] = f(x;\phi) \tag{2}$$

$$q(z|x) = N(z;\tilde{\mu},\tilde{\sigma}^2 I) \tag{3}$$

Using stochastic gradient variational Bayes estimator and with proper re-parameterization, the parameters can be adjusted by maximizing ELBO with the following criterion:

$$D_{KL}(q(c|x)||p(c|z)) \equiv 0 \tag{4}$$

where $p(c|z)$ is the prior for specific mixture c. In this work, we utilize VaDE to derive our acoustic network representation and then further encode frame-level LLDs to the latent representation $\bar{x}_l$ using the learned encoding network f.

Table 1: *A summary on Spearman correlation obtained on valence and activation of regression experiments in the NNIME and CIT database. All results have p-value* $< 10^{-3}$. *The top half part shows the results of Exp I as baseline comparison and the lower half part shows the results of augmentation and different fine-tune strategies. A general VaDE representation is denoted as* R_G *and a dyad-specific VaDE representation is* R_D. R_A *is the augmented representation concatenated by* R_G *and* R_D.

	NNIME						CIT					
R_G	E	P	E_{AG}	E_{VaDE}	P_{AG}	P_{VaDE}	E	P	E_{AG}	E_{VaDE}	P_{AG}	P_{VaDE}
Act.	0.603	0.635	0.611	**0.679**	0.627	**0.670**	0.417	0.424	0.365	0.452	0.486	**0.521**
Val.	0.246	0.571	0.309	0.365	0.465	**0.580**	0.326	0.322	0.309	0.325	0.357	**0.372**
	R_G		R_D			R_A			R_D			R_A
	P_{VaDE}	fine-tune	adapt	frozen	fine-tune	adapt	frozen	P_{VaDE}	fine-tune	adapt	frozen	fine-tune	adapt	frozen
Act.	0.670	0.408	0.303	0.531	0.639	**0.700**	0.696	0.521	0.137	0.406	0.208	0.511	0.540	**0.544**
Val.	0.580	0.355	0.213	0.290	0.596	0.598	**0.604**	0.372	0.185	0.257	0.234	0.318	0.386	**0.387**

2.2.3. Dyad-Augmented Deep Embedding Networks

In this work, we learn two VaDE networks, i.e., a general VaDE network and a dyad-specific VaDE network. The general VaDE with encoding network $f(x;\phi)$ and decoding network $h(z;\theta)$ is learned from the entire corpus that describes general expressive vocal behaviors. The dyad-specific VaDE with encoding network $f'(x';\phi')$ and decoding network $h'(z';\theta')$ is used to represent the unique interaction dynamics for a specific dyad-pair. Since the available data for a specific dyad-pair is often limited, a robust dyad-specific VaDE is learned by performing fine-tuning on the general VaDE network.

Conventional fine-tuning often suffers from the problem of forgetting effect, i.e., the network 'forgets' its modeling power in the original domain after fine-tuning [23, 24]. In terms of dyad-pair behavior dynamics, this forgetting can be detrimental. Intuitively, the dyad-specific dynamics should not deviate too drastically from general behavior dynamics but to add fine-grained auxiliary dynamics. Hence, we propose to learn the dyad-specific network using a *frozen* transferring technique (similar to [25]), which freezes the general VaDE encoder-decoder network weights except updating the middle latent layer jointly with GMM prior by fine-tuning using the data of the specified dyad-pair.

In summary, for each individual, we encode the frame-level LLDs using both the encoding networks, (f, f'), of general VaDE and dyad-specific VaDE model to derive our dyad-augmented deep embedding latent vectors.

2.2.4. Dialog-level Emotion Recognition

Since each dialog is different in its duration, the encoded LLDs into our dyad-augmented VaDE networks would result in a varying number of representation sequences. We compute the gradient log-likelihood function, i.e., Fisher scoring (indicating the direction of λ to better fit $\bar{x}_l$), with respect to the first and second order statistics of the learned latent VaDE-GMM parameters to further encode a sequence of acoustic latent representation $\bar{x}_l$ into a fixed-length representation (also terms as GMM-based Fisher-vector encoding [26]). The use of Fisher-vector encoding has been shown to be competitive in speech-related tasks of paralinguistic recognition [27], presentation scoring [28], and emotion recognition [29, 30]. The dialog-level acoustic vectors that integrates both the general representation and the dyad-specific dynamics is derived by concatenating the general Fisher-scoring vector with the dyad-specific Fisher-scoring vector. This is the final feature vector input that is used to train a support vector regression for dialog-level emotion recognition.

3. Experimental Setup and Result

In this work, we compare our dyad-augmented VaDE acoustic representation with different models in tasks of activation and valence regression on the two databases. Two different analyses experiments are carried out:

- **Exp I** : Comparison to other vocal representations for dialog-level emotion recognition
- **Exp II**: Comparison between different dyad-specific augmented representation techniques

Exp I is carried out by comparing with multiple representation learning network schemes with two different low level feature sets when using only the general representation. Exp II is designed to examine the effect of dyad-augmented representation on emotion recognition task, which is derived from both the general network embedding and dyad-specific network embedding and further compare to other fine-tuning strategies.

All the experiments are carried out using leave-one-dyad-out cross validation, the support vector regression with linear kernel and fixed parameters ($C = 1$), and Spearman correlation as the evaluation metric. Table 2 lists the network parameters for general VaDE model and dyad-specific VaDE model.

3.1. Exp I: Comparison to Other Representations

A list of feature representations to compare is shown below:

- $\mathbf{E}$: Fisher scoring representation on eGeMAPS low-level descriptors in 20ms frame size and 10ms step size computed using opensmile [31]
- $\mathbf{E_{AG}}$: Fisher scoring representation derived from GMM separately trained on the autoencoder latent layer using low-level descriptors of eGeMAPs
- $\mathbf{E_{VaDE}}$: Fisher scoring representation derived from the VaDE latent layer using low-level descriptors of eGeMAPs
- $\mathbf{P}$: Fisher scoring representation on low-level acoustic features computed using Praat described in section 2.2.1
- $\mathbf{P_{AG}}$: Fisher scoring representation derived from GMM separately trained on the autoencoder latent layer using low-level descriptors from Praat
- $\mathbf{P_{VaDE}}$: Fisher scoring representation derived from the VaDE latent layer joint GMM prior using low-level descriptors from Praat

The results of Exp I are demonstrated in the top half of Table 1. We show the recognition results obtained using each feature set (E: eGeMAPS, P: Praat), autoencoder-based methods (E_{AG} or P_{AG}), and the VaDE methods (E_{VaDE} and P_{VaDE}).

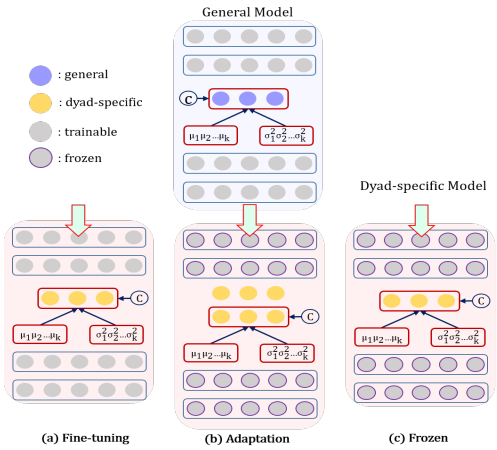

Figure 2: *It shows a schematic of different fine-tuning strategies: (a) a simple fine-tuning method that uses low learning rate and few epochs, (b) an adaptation with an additional layer for capacity expansion, and (c) our proposed fine-tuning approach with trainable latent layer while the others frozen.*

Our experiments show that the best recognition accuracies obtained using general model only is by learning the representation using the VaDE approach, which achieves correlations of 0.670 and 0.580 in NNIME database on activation and valence dimension, and 0.521 and 0.372 on activation and valence dimension respectively in CIT database.

We observe that the VaDE representation learned using both feature sets achieve the best performance compared to the other learning schemes. It is also interesting to see that E_{AG} or P_{AG}, which learns the autoencoder and GMM separately, can sometimes exceed the accuracy obtained from E or P though never surpass the VaDE model (jointly optimized the autoencoder with GMM parameters). It seems to indicate that the advanced nonlinear representation modeling power in these network structures benefits from the loosely-regularized GMM in their latent layer when learning from data.

3.2. Exp II: Comparison between Dyad-Augmentations

Our proposed dyad-augmented vocal representation (R_A) includes a general VaDE representation (R_G) and a dyad-specific VaDE (R_D). It obtains the best correlation, i.e., 0.700 and 0.604 for activation and valence on the NNIME, and 0.544 and 0.387 for activation and valence on the CIT. This result shows a relative gain of 4.48% and 4.14% compared with R_G using P_{VaDE} (without dyad-augmentation) in recognizing activation and valence in the NNIME database. Similar improvement in obtaining a 4.41% and 4.03% relative performance on activation and valence also holds in the CIT database (Table 1). We observe that using R_D only would negatively impact the recognition correlation due to its inadequate modeling power on acoustic representation (contains only the dyad-specific variability from limited data samples). These results shows that our augmentation technique which incorporate dyadic-specific unique behavioral dynamics is important in improving the emotion recognition for an individual in an face-to-face setting.

Since the dyad-specific VaDE is learned by fine-tuning on general VaDE. We further analyze and compare various fine-tuning techniques in deriving the dyad-specific VaDEs. We examine the three widely-used fine-tuning methods illustrated in Figure 2. First, we simply fine-tune the network using the dyad-specific data. Second, we add an additional hidden layer and make the latent layers trainable with the other original layers frozen (denoted as adaptation in Figure 2). Third is our proposed method that makes the latent layer trainable with the original encoder-decoder layers frozen (denoted as frozen strategy in Figure 2). All of these fine-tune strategies are carried out using small learning rate and few epochs.

Table 2: *The VaDE architectures and configurations used in our emotion recognition experiments.*

Parameters		*minibatch*	*epoch*	C_{GMM}	*lr*
CIT	$\mathbf{R_G}$	100	50	16	0.0002
	$\mathbf{R_D}$	100	10	8	0.00002
NNIME	$\mathbf{R_G}$	20	50	16	0.002
	$\mathbf{R_D}$	20	10	16	0.00002

Table 1 bottom half summarizes results obtained from different fine-tuning strategies. We observe that techniques based on adaptation and frozen methods can outperform typical fine-tuning method. The frozen strategy obtains the best performance among the three. Generally, methods with frozen strategy is favorable for the augmented representation compared to the adaptation method, which may due to the fact that the adaptation adds an non-initialized layers and may be too complex to model the 'additional' intricate vocal interaction dynamics between the interlocutors.

4. Conclusion

In this paper, we propose a novel framework in learning a dyad-augmented deep variational vocal representations that integrates the unique dyadic interaction dynamics to improve individual's dialog-level emotion recognition. By encoding and concatenating an individual acoustic features resulting from using two deep generative networks, a general and dyad-specific VaDE network, we achieve an improved dialog-level emotion recognition accuracies on activation and valence dimensions demonstrated on two different corpora. The dyad-specific VaDE representation is learned through fine-tuning general network using a frozen strategy. Our analyses further demonstrate that such a frozen fine-tuning technique is important in obtaining the improved accuracy. To the best of our knowledge, this is one of the first works in embedding the natural dyadic behavior dynamics directly at the level of acoustic representation in task of speech emotion recognition.

There are multiple future directions. One of the immediate work is to include multimodal behavior information, e.g., body movement and gestural information, to achieve a further improved emotion recognition by leveraging the mutual-dependency across behavior modalities and further between interacting dyads. The representation learning framework offers flexibility in sophisticated behavior and even interaction pattern modeling at the low-level descriptors level. We will further validate and advance upon the VaDE-based behavior representation framework on an expanded list of dyadic interaction databases. By continuing to develop algorithm in achieving a robust emotion recognition system would contribute to the enabling of the next generation applications in not only human-centered research and development but also create a tangible impact on mental health-related applications [32].

5. References

[1] D. Davis, "Determinants of responsiveness in dyadic interaction," in *Personality, roles, and social behavior*. Springer, 1982, pp. 85–139.

[2] R. S. Lazarus, "Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping," *Journal of Personality*, vol. 74, no. 1, pp. 9–46, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1467-6494.2005.00368.x

[3] A. Jakkam and C. Busso, "A multimodal analysis of synchrony during dyadic interaction using a metric based on sequential pattern mining," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6085–6089.

[4] D. del Valle-Agudo, J. Calle-Gómez, D. Cuadra-Fernández, and J. Rivero-Espinosa, "Interpretation and generation incremental management in natural interaction systems," *Interacting with Computers*, vol. 24, no. 2, pp. 78–90, 2012.

[5] V. Gallese, "The manifold nature of interpersonal relations: the quest for a common mechanism," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 358, no. 1431, pp. 517–528, 2003. [Online]. Available: http://rstb.royalsocietypublishing.org/content/358/1431/517

[6] F. Alam, S. A. Chowdhury, M. Danieli, and G. Riccardi, "How interlocutors coordinate with each other within emotional segments?" in *COLING*, 2016.

[7] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.

[8] M. Berking and P. Wupperman, "Emotion regulation and mental health: recent findings, current challenges, and future directions," *Current opinion in psychiatry*, vol. 25, no. 2, pp. 128–134, 2012.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ser. ICMI '04. New York, NY, USA: ACM, 2004, pp. 205–211. [Online]. Available: http://doi.acm.org/10.1145/1027933.1027968

[10] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Computers in Human Behavior*, vol. 65, pp. 267–275, 2016.

[11] M. Liu, D. Fan, X. Zhang, and X. Gong, "Human emotion recognition based on galvanic skin response signal feature selection and svm," in *Smart City and Systems Engineering (ICSCSE), International Conference on*. IEEE, 2016, pp. 157–160.

[12] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.

[13] Z. Yang and S. Narayanan, "Analyzing temporal dynamics of dyadic synchrony in affective interactions," *Interspeech 2016*, pp. 42–46, 2016.

[14] Z. Yang, B. Gong, and S. Narayanan, "Weighted geodesic flow kernel for interpersonal mutual influence modeling and emotion recognition in dyadic interactions," in *In Proceedings of Seventh International Conference on Affective Computing and Intelligent Interaction*, October 2017.

[15] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2401–2404.

[16] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April 2012.

[17] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April 2013.

[18] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions." in *Interspeech*, 2009, pp. 1983–1986.

[19] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1965–1972.

[20] A. Metallinou, Z. Yang, C.-c. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Language resources and evaluation*, vol. 50, no. 3, pp. 497–521, 2016.

[21] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "Nnime: The nthu-ntua chinese interactive multimodal emotion corpus," in *In Proceedings of Seventh International Conference on Affective Computing and Intelligent Interaction*, October 2017.

[22] M. D. Hoffman and M. J. Johnson, "Elbo surgery: yet another way to carve up the variational evidence lower bound," in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[23] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetful learning for domain expansion in deep neural networks," *arXiv preprint arXiv:1711.05959*, 2017.

[24] Y.-X. Wang, D. Ramanan, and M. Hebert, "Growing a brain: Fine-tuning by increasing model capacity," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *European Conference on Computer Vision*. Springer, 2016, pp. 435–442.

[26] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[27] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[28] S. W. Hsiao, H. C. Sun, M. C. Hsieh, M. H. Tsai, Y. Tsao, and C. C. Lee, "Toward automating oral presentation scoring during principal certification program using audio-video low-level behavior profiles," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[29] W. C. Lin and C. C. Lee, "Computational analyses of thin-sliced behavior segments in session-level affect perception," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2018.

[30] 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)*, March 2017, pp. 5820–5824.

[31] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[32] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.