INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Use of Agreement/Disagreement classification in Dyadic Interactions for Continuous Emotion Recognition

Hossein Khaki, Engin Erzin

Multimedia, Vision and Graphics Lab,
Koç University, Istanbul, Turkey
hkhaki13,eerzin@ku.edu.tr

Abstract

Natural and affective handshakes of two participants define the course of dyadic interaction. Affective states of the participants are expected to be correlated with the nature or type of the dyadic interaction. In this study, we investigate relationship between affective attributes and nature of dyadic interaction. In this investigation we use the JESTKOD database, which consists of speech and full-body motion capture data recordings for dyadic interactions under agreement and disagreement scenarios. The dataset also has affective annotations in activation, valence and dominance (AVD) attributes. We pose the continuous affect recognition problem under agreement and disagreement scenarios of dyadic interactions. We define a statistical mapping using the support vector regression (SVR) from speech and motion modalities to affective attributes with and without the dyadic interaction type (DIT) information. We observe an improvement in estimation of the valence attribute when the DIT is available. Furthermore this improvement sustains even we estimate the DIT from the speech and motion modalities of the dyadic interaction.

Index Terms: Multimodal Continuous Emotion Recognition, human-computer interaction, Dyadic Interaction Type.

1. Introduction

Social signals are perceivable stimuli that, either directly or indirectly, convey information concerning social actions, social interaction, attitudes, social emotions and social relations [1]. Through social signals of agreement and disagreement in a communicative interaction participants can share convergent or divergent opinions, proposals, goals, attitudes and feelings. In recent literature common types of such social interaction are the group meeting scenarios [2–5], political debates [6–8], theatrical improvisations [9] and broadcast conversations [10, 11].

One of the main parameters in a general discussion is the type of the discussion, which can control the level of emotion of the participants in a dyadic interaction. Based on psychological models, Activation-Valance-Dominance (AVD) space, which describes the intensity, level of pleasure, and amount of control of the emotion, is one of the continuous model to describe the affective state [12].

In this paper, we investigate the relationship between affective attributes and nature of dyadic interaction and analyze the performance of an AVD recognition system over agreement and disagreement types. Particularly, we propose a DIT-based continuous emotion recognition system (DIT-CER) that initially classifies the discussion into agreement or disagreement classes and then recognizes the AVD. To address this issue, we use the JESTKOD database, which consists of speech and full-body motion capture data recordings for dyadic interactions under agreement and disagreement scenarios [13,14]. The dataset also has affective annotations in activation, valence and dominance (AVD) space. Our experimental results indicate that the valence attribute recognition improves when DIT is available.

The remainder of the paper is organized as follows. We review the current datasets for emotion recognition and describe the JESTKOD dataset in Section 2. Moreover, we provide statistical analysis of annotation in Subsection 2.2.1 to illustrate the relationship between AVD and DIT. Then, in section 3, we present the details of the DIT-CER system. We then demonstrate the performance of the proposed system based on experimental results in Section 4. Finally, we conclude the paper and outline future study directions in Section 5.

2. Multimodal Dyadic Interaction Database

2.1. Literature Review

A variety of multimodal databases, containing continuous affect annotations, are publicly available for research purposes. One such a database is the HUMAINE database, which includes a large collection of multimodal naturalistic and induced recordings [15]. Another one is the SEMAINE database, consisting of audio-visual data in the form of conversations between participants and a number of virtual characters with particular personalities [16]. The acted audio-visual MSP-IMPROV database investigates emotional behaviors during conversational dyadic improvisations [17]. Technical setups, scenarios and challenges in building a motion capture database for virtual human animation are explored in [18]. An interactive emotional dyadic motion capture, named the USC IEMOCAP database, is presented in [19], which is a multimodal and multi-speaker database of improvised and scripted dyadic interactions. The USC CreativeIT database contains full-body motion capture information in the context of expressive theatrical improvisations [9, 20]. The database is annotated using the valence, activation and dominance attributes, as well as the theater performance ratings such as interest and naturalness. However, speech and body gestures are rather amplified and pretentious in this database since interaction performances are theatrical.

2.2. JESTKOD Database

Our main motivation to construct the JESTKOD database is to address more natural and affective dyadic interactions, providing a valuable asset to investigate gesticulation during spoken interactions. The JESTKOD database consists of dyadic interaction recordings of 10 participants, 4 female and 6 male, ages from 20 to 25. Agreement and disagreement interactions of the

http://dx.doi.org/10.21437/Interspeech.2016-407

5 dyads are collected in 5 sessions, all in Turkish. Each participant interacted with the same partner for both agreement and disagreement settings and only appeared in one session. In each session, there are 9-13 clip recordings of 2-4 minutes, where participants pick a topic that they agree or disagree and engage into a dyadic interaction. The total duration of the recordings is 259 minutes. Recordings are performed by a high-definition video camera, full body motion capture system and Sony condenser tie-pin microphones. We used the OptiTrack Flex 13 [21] system and the *Motive* software [22] for the full body motion capture, which consists of 12 infrared cameras capturing 21 body joints at 120 fps with a resolution of 1280x1024.

Topics of the dyadic interactions are set by the moderator of the session using a preliminary information form, where participants are asked to state their favorite and disliked such as soccer teams, foods, restaurants, world cuisines, computer games, movies, operating systems, game consoles. Using these forms, we compiled a list of topics and paired up the participants with proper topics to create agreement/disagreement interactions during the recordings. In the JESTKOD database, we have 55 and 43 dyadic interactions in agreement and disagreement with total durations of 154 and 105 minutes, respectively.

Annotation effort is carried over for each participant in the recordings and for each dimensional attribute separately. A joystick interface is used with the *GTrace* software to deliver continuous-time annotations of the activation, valence and dominance attributes. Annotations per attribute are performed by the same annotator. A total of six annotators contribute to collect three sets of annotations for valence and dominance, and four sets of annotations for activation attribute. The mean correlation values are reported in Table 1 for the AVD attributes.

Table 1: Mean Pearson's correlation between the ground truth and individual annotations for activation, valence and dominance attributes under different interaction types

	Activation	Valence	Dominance
Agreement	0.5456	0.5975	0.7562
Disagreement	0.5680	0.5300	0.7176
All	0.5568	0.5638	0.7369

2.2.1. *Statistical Analysis of Annotations*

We investigate effect of agreement and disagreement interaction scenarios on the AVD annotations. For this aim, we compute histograms of the ground truth annotations for each attribute under agreement and disagreement scenarios. The histograms of activation, valence and dominance are depicted in Figure 1. Activation and dominance attributes do not convey significant distribution differences for agreement and disagreement interaction scenarios. However, histograms of the valence attribute differ significantly for agreement and disagreement, where the histogram for agreement and disagreement is shifted to the positive and negative valence axis, respectively. This behavior is expected for the valence attribute under agreement (positive) and disagreement (negative) interaction scenarios. Hence, we can argue that given the agreement/disagreement classification information the estimation of valence should improve. This issue is the main question of the current study.

Furthermore, to quantify statistical difference between agreement and disagreement distributions, We utilized the Kullback-Leibler divergence (KLD). We can define KLD and symmetric KLD respectively as,

$$D_{KL}(P_A||P_D) = \sum_k P_A(k) \log \frac{P_A(k)}{P_D(k)}, \quad (1)$$

and

$$D_{KL}(P_A, P_D) = \frac{1}{2}(D_{KL}(P_A||P_D) + D_{KL}(P_D||P_A)), \quad (2)$$

where $P_A()$ and $P_D()$ are respectively probability distributions over agreement and disagreement scenarios, and k runs over the sample space of activation, valence or dominance. Table 2 presents the symmetric KLD distances for the activation, valence and dominance attributes. Agreement and disagreement distributions have the strongest difference for the valence with a KLD distance of 1.1844. Activation attribute seems to yield no significant statistical difference for the agreement and disagreement, where as dominance presents a small KLD distance of 0.2079.

Based on the KLD, we conclude that the estimation of valence can improve given the agreement disagreement information and we don't expect any significant gain for activation and valence. We analyze this claim by proposing a simple separated classifier for two different dyadic interaction type.

Table 2: KLD distances between agreement/disagreement interactions for the activation, valence and dominance attributes

$D_{KL}(P_A\|\|P_D)$		
Activation	Valence	Dominance
0.0327	1.1844	0.2079

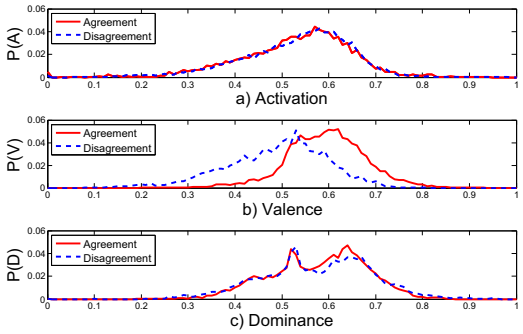

Figure 1: Histograms of the activation, valence and dominance attributes under agreement and disagreement interactions

3. Continuous emotion recognition using DIT estimation

In this part, first we describe the DIT estimator and then explain the CER based on DIT estimator which we call it DIT-CER system.

3.1. Dyadic interaction type estimation

In our previous study [23], we investigate a multimodal two-class dyadic interaction type (DIT) estimation approach of agreement and disagreement classes from speech and motion

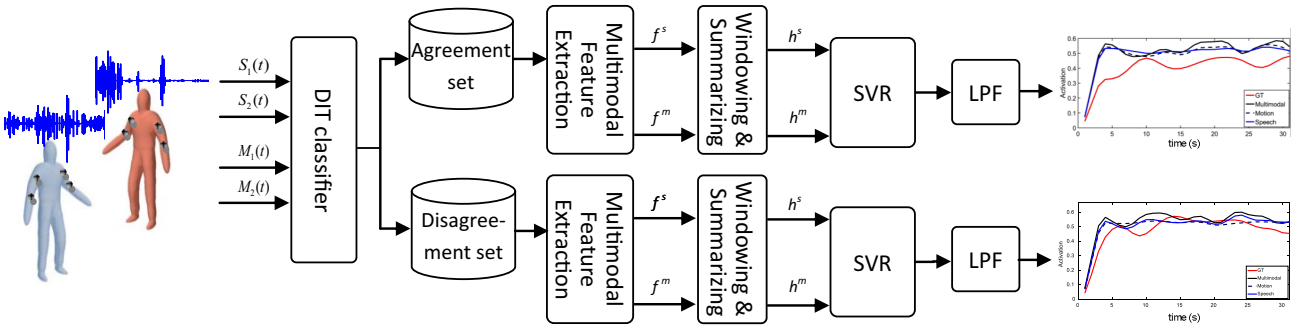

Figure 2: Dyadic Interaction Type - Continuous Emotion Recognition (DIT-CER) system.

modalities. In that model, we utilize the 13 dimensional MFCC feature vector together with its first and second order derivatives for speech modality and the Euler rotation angles in directions (x,y,z) of the arm and forearm joints together with their first derivatives for motion modality. We apply i-vector and statistical functions to convert the matrix of features to a summarized feature vector by joint speaker model (JSM), i.e., summarizing the features of two participants together, or split speaker model (SSM), i.e., summarizing the features of each participants separately. According to our experimental evaluations, the accuracy of agreement and disagreement classification exceeds 80% when we have more than 15 sec of data in the multimodal scenario. Moreover, the performance of JSM with i-vector and SSM with statistical function is appropriate. In this paper, we use the SSM with statistical function summarizer unit to divide the dataset into agreement and disagreement sets and study the relationship between the DIT and the affective state.

3.2. Continuous emotion recognition

Our DIT based continuous emotion recognition (DIT-CER) system is based on the DIT estimator explained in Section 3.1 and window level support vector regression (SVR) method, depicted in figure 2. Firstly, we utilize the DIT estimator to divide the dataset into two set of agreement and disagreement sets.Then, for each set, we extract the speech, f^S, and motion, f^M, feature vectors. We use same mfcc and Euler rotation angles features as in Section 3.1. We select the acoustic frame size such that we attain same number of frame as motion capture system.

Since the affective state varies slowly, we utilize a window level framework as [24] and collect frame level feature vectors over the temporal duration of the window and construct matrices of features. We fill the silence frames for the speech modality by random noise with normal distribution and construct speech feature matrix as $F_k^S = [f_1^S \cdots f_N^S]$ for the k-th window with dimensions $39 \times N$. Similarly, the motion feature matrix is constructed as $F_k^M = [f_1^M \cdots f_N^M]$ with dimensions $24 \times N$ without silence replacement. Then, we summarize the features over overlapping temporal windows using a summarization function, $F : R^{m*N} R^k$, in which m is the dimension of speech or motion features, N is the number of frames over a time window, and k is the dimension of the summarized features. Based on [24], we utilize a variety of statistical functions such as mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, the lower and upper quantiles (corresponding to the 25th and 75th percentiles) and the interquantile range followed by PCA to reduce the dimension, to extract the summarized feature vector, h_i^S and h_i^M for the speech and motion modalities, respectively, where i is the window index. In addition, we set the mean value over the temporal window of each emotional attributes a_i as the corresponding annotation. We apply Support Vector Regression (SVR) to map the speech, motion, or multimodal feature spaces to activation, valence, and dominance (AVD) domain. Since motion and speech modalities have the same frame rate in feature level, we simply perform feature fusion for multimodal case.

Table 3: Notations and descriptions of the test conditions

Test name	Estimated parameter	Test details
Speech	$\hat{a}^S$	Using speech over vocal parts
Motion	$\hat{a}^M$	Using motion modality
Multimodal	$\hat{a}^{SM}$	Using speech and motion, non-vocal speech frames are replaced by random noise
Vocal Multimodal	$\hat{a}^{VSM}$	Using speech and motion over vocal parts

After the SVR, the estimated attributes, $\hat{a}_i$, are low-pass filtered for smoothing as in [24]. Although the mean square error between the estimated and mean of the annotators AVD dimensions is a possible evaluation metric, we choose to evaluate the performance with the correlation metric since variation of the AVD dimensions is more important than the exact values.

4. Experimental evaluations

In the JESTKOD database, we have 5 different sessions. We perform the evaluation in a leave-one-session-out training, where we test the clips of a session at a time and train models on the remaining recordings. Hence, our test is a speaker independent evaluation. We employ an automatic Voice Activity Detector (VAD) [25] to replace the silent segments of the speech recordings by random noise with normal distribution. An acoustic frame is computed every 8.33 msec over 16.66 msec analysis windows to attain the same frame rate as motion capture, which is 120 fps. Each acoustic frame is a 39 dimensional vector, which includes the Energy, the first 12 Mel Frequency Cesptral Coefficients plus the first and second time derivatives. For the motion modality features, we utilize the Euler rotation angles in directions (x,y,z) of the arm and forearm joints together with their first derivatives over each frame. For DIT estimation, we generate a summarized feature vector per

Table 4: Mean correlation percentage of estimated and ground truth activation, valence, and dominance attributes using speech, motion, and multimodal cues.

	Activation				Valence				Dominance			
Dataset	**S**	**M**	**SM**	**VSM**	**S**	**M**	**SM**	**VSM**	**S**	**M**	**SM**	**VSM**
Agreement clips	52.45	38.30	47.95	54.28	61.64	49.80	54.30	62.65	55.87	46.25	49.08	54.50
Disagreement clips	50.75	33.99	41.41	52.24	44.77	30.07	36.92	47.38	58.98	49.81	50.45	55.68
DIT-CER	51.72	36.46	**45.16**	53.41	**54.43**	**41.38**	**46.87**	56.13	57.20	47.77	**49.66**	55.01
EDIT-CER	**52.30**	**37.37**	44.43	53.44	54.05	38.29	46.29	**56.45**	56.80	46.62	49.08	54.85
CER w/o DIT	51.70	37.30	45.11	**53.50**	52.44	39.03	44.31	55.95	**57.85**	**48.46**	49.31	**56.34**

clip and classify it to agreement and disagreement sets. Then, we perform the CER by rate of 750 ms over the 1.5 s overlapped windows and generate the summarized feature vector per window. We use the statistical function to summarize the matrix of features and adjust the PCA output dimension to preserve 90% of the total variance for the output of statistical function. In the estimation of affective attributes we use radial basis function kernel SVR from the LibSVM package [26].

We define four different tests, speech-only, motion-only and two multimodal conditions. Their summary and notation are given in Table 3. In test conditions M and SM, affective attributes are estimated over all windows. On the other hand in test conditions S and VSM, training and estimation are performed over the vocal windows only. Hence, a fair comparison can be done between the M and SM or S and VSM.

Evaluations are performed over different sets of data to realize the effect of interaction type. In the evaluations, we calculate the mean correlation across recordings between the affective state, a_i, and the estimated affective state $\hat{a}_i$. We used the true interaction type in Agreement clips, Disagreement clips, and DIT-CER, the estimated interaction type in EDIT-CER, and no interaction information in CER w/o DIT, as listed in Table 4, respectively. We have three main columns for Activation-Dominance-Valence in Table 4. In each main column, we have four columns corresponding to the unimodal and multimodal test conditions. Note that the presented results in DIT-CER set are equal to the weighted mean of the Agreement/Disagreement clips results. The weights correspond to the number of clips in Agreement/Disagreement sets.

The performance of CER with true or estimated DIT is almost always higher than CER without DIT for valence recognition. However, this is not true for activation and dominance. The attained improvement for valence recognition and reaching the same performance for activation and dominance were expected based on the statistical analysis of data in Section 2.2.1.

Using multimodal information improves the performance significantly. This improvement can be seen by comparing the motion (M) and multimodal (SM) columns of each emotional attributes in Table 4. On the other hand, by comparing the speech (S) and vocal multimodal (VSM) columns, adding motion modality to speech modality improves the activation and valence recognition performance but decreases the performance of dominance.

By comparing the multimodal and vocal multimodal results, third and fourth columns, we interpret that the vocal parts have higher mean correlation for all AVD dimensions. Finally, by comparing the GT in Table 1 and the highest mean correlation over each main columns in Table 4, we observe marginal difference.

5. Conclusions and future work

In this paper, we investigated the relationship between the emotional status, Activation-Dominance-Valence (AVD), and dyadic interaction type (DIT). In this investigation, we used the JESTKOD database, which consists of speech and full-body motion capture data recordings for dyadic interactions under agreement and disagreement scenarios followed by AVD annotation. Initially, we analyzed the annotation under agreement/disagreement conditions and observed significant difference between the two conditions for valence dimension by calculating the symmetric Kullback-Leibler divergence metric. We then proposed a DIT-based continuous emotion recognition system (DIT-CER), which initially separates the data into agreement and disagreement sets, then estimates the emotion for each set separately. We demonstrated the continuous AVD estimation performance under speech, motion, multimodal and vocal multimodal tests and realized that given accurate or estimated DIT information, the estimation of valence improves and the estimation of other two attributes changes marginally.

Next step to discover the relation between the AVD and DIT would be investigating other separation-fusion methods to build a more robust system to the division of the data into different classes. Moreover, the agreement/disagreement classification can be defined over the utterance level, which is shorter than clips, instead of clip level and the system may gain more from the higher resolution provided by utterance level segmentation.

6. Acknowledgements

This work was supported by TÜBİTAK under Grant Number 113E102.

7. References

[1] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, Jan 2012.

[2] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10579-007-9040-x

[3] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 305–317, March 2005.

[4] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceed-*

ings of HLT-NAACL 2003–short Papers - Volume 2, ser. NAACL-Short '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 34–36.

[5] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ser. ACL '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[6] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 5089–5092.

[7] K. Bousmalis, L. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, March 2011, pp. 746–752.

[8] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: a database of political debates for analysis of social interactions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, ser. ACII '09, 2009.

[9] A. Metallinou, Z. Yang, C.-c. Lee, C. Busso, S. Carnicke, and S. 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*, 2015.

[10] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, June 2008, pp. 865–868.

[11] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5556–5559.

[12] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[13] E. Bozkurt, H. Khaki, S. Kececi, B. B. Turker, Y. Yemez, and E. Erzin, "Jestkod database: Dyadic interaction analysis," in *Signal Processing and Communications Applications Conference (SIU), 2015 23th*. IEEE, 2015, pp. 1374–1377.

[14] S. Kececi, E. Erzin, and Y. Yemez, "Analysis of jestkod database using affective state annotations," in *2016 24nd Signal Processing and Communications Applications Conference (SIU)*.

[15] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, A. Paiva, R. Prada, and R. Picard, Eds. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 488–500.

[16] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, Jan 2012.

[17] C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[18] A. Heloir, M. Neff, and M. Kipp, "Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization," in *Proc. of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010.

[19] 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, pp. 335–359, Nov. 2008.

[20] A. Metallinou, C. C. Lee, C. Busso, S. Carnicke, and S. S. Narayanan, "The USC CreativeIT Database : A Multimodal Database of Theatrical Improvisation," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, May 2010.

[21] OptiTrack, "Flex 13 system," 2016. [Online]. Available: http://www.optitrack.com/products/flex-13/

[22] ——, "Motive - Optical motion capture software," 2016. [Online]. Available: http://www.optitrack.com/products/motive/

[23] H. Khaki, E. Bozkurt, and E. Erzin, "Agreement and disagreement classification of dyadic interactions using vocal and gestural cues," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2762–2766.

[24] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[25] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 1997.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.