INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Cross-Subject Continuous Emotion Recognition using Speech and Body Motion in Dyadic Interactions

Syeda Narjis Fatima and Engin Erzin

Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, Istanbul, Turkey
[sfatima13,eerzin]@ku.edu.tr

Abstract

Dyadic interactions encapsulate rich emotional exchange between interlocutors suggesting a multimodal, cross-speaker and cross-dimensional continuous emotion dependency. This study explores the dynamic inter-attribute emotional dependency at the cross-subject level with implications to continuous emotion recognition based on speech and body motion cues. We propose a novel two-stage Gaussian Mixture Model mapping framework for the continuous emotion recognition problem. In the first stage, we perform continuous emotion recognition (CER) of both speakers from speech and body motion modalities to estimate activation, valence and dominance (AVD) attributes. In the second stage, we improve the first stage estimates by performing CER of the selected speaker using her/his speech and body motion modalities as well as using the estimated affective attribute(s) of the other speaker. Our experimental evaluations indicate that the second stage, cross-subject continuous emotion recognition (CSCER), provides complementary information to recognize the affective state, and delivers promising improvements for the continuous emotion recognition problem.

Index Terms: Continuous emotion recognition, dyadic emotion estimator, side emotional information, cross-subject continuous emotion recognition (CSCER), Gaussian mixture regression, Activation, Valence, Dominance

1. Introduction

Affective human interactions are well defined psychological paradigms that reflect a multitude of human behaviors [1, 2]. Emotion is an inextricable constituent of affective interactions. It can be described by three primitive dimensional attributes: activation, valence and dominance [3, 4]. Activation captures the intensity of emotional experience, valence describes the positive and negative levels of pleasure related to an emotion while dominance describes the level of control of a person during an emotional experience [5]. Collectively, the three attributes summarize the global emotion state of the participant in an interaction. Understanding emotion is crucial in development of realistic emotionally intelligent machines, better emotion prediction in speech synthesis, and tracking of emotional responses in social scenarios such as dyadic interactions and counseling.

Continuous emotion recognition (CER) has increasingly attracted interest over the past few years [5, 6]. Speech and motion modality information are extensively used for continuous emotion recognition [7, 8, 9] based on a wide variety of techniques such as k-means clustering, support vector regression, neural networks and Gaussian mixture model (GMM) regression [10, 11, 12, 13, 14]. The prevalent challenges in CER to date necessitate improved techniques because the estimated performance is context-dependent and varies between databases. This shortcoming may be mitigated if the overall relationship between the participant of a conversation is taken in account during the process of emotion estimation.

Dyadic interactions are reflective of a complex interplay of emotions reflected through speech and body movements where the dynamic interplay in the directionality of influences can be in synchrony, diverging or accommodating nature [13, 15, 16, 17, 18]. The dependence between discrete emotional states of the dialog partners is established in [19]. Further, in [20], the relationship of activation and valence with gesture and speech based synchrony is explored. These studies suggest that there exits a relationship between the emotional attributes of the participants in a dyadic interaction. We hypothesize that while the three dimensions are independent, there exists a synchrony between emotions at a cross-subject level. This dependency should reflect itself in the prediction of emotion in a dyadic setting. Such a cross-speaker emotional relationship may be very important in continuous emotion recognition for participants of an interaction, and to date, no such work is reported in literature.

Therefore, in this work we hypothesize that utilizing the cross-subject, cross-emotion relationship between the three different emotional attributes of participants involved in a dyadic interaction can improve CER. We propose a novel two-stage framework for emotion estimation utilizing the cross-subject dependence at the emotion-level. In the first stage, we perform conventional CER for participants in a dyadic interaction using their speech and body motion data to estimate activation, valence and dominance (AVD) attributes independently. We use Gaussian Mixture Model (GMM) regression for attribute estimation similar to [10, 11, 21]. In the second stage, we improve the first stage estimates by performing CER of the selected speaker using her/his speech and body motion modalities as well as using the estimated affective attribute(s) of the other speaker as side emotional information (SEI). Our experimental evaluations indicate that the two stage, cross-subject continuous emotion recognition (CSCER), provides complementary information to recognize the affective state, and delivers promising improvements for the continuous emotion estimation problem. Our work shows that cross-speaker dependencies can be parameterized to achieve realistic cross-speaker emotion estimation in a dyadic setting. To the best of our knowledge, this work is the first of its kind towards the realization of a dyadic emotion estimator utilizing cross-subject emotional dependencies.

The remainder of the paper is organized as follows. Section 2 describes the database, feature extraction and summarization followed by conceptualization of side emotional information for CER. Detailed description of our two-stage CSCER framework is also presented in Section 2. Experimental evaluations and results are presented in Section 3. Finally, Section 4 gives a summary of our current work and provides some further directions.

http://dx.doi.org/10.21437/Interspeech.2017-1413

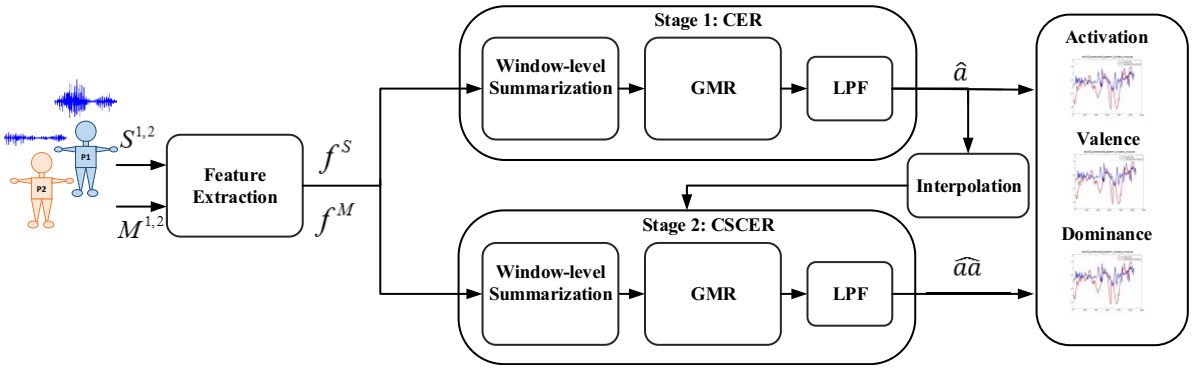


Figure 1: *Cross-subject continuous emotion recognition (CSCER) framework.*

2. Methodology

The block diagram of the proposed two stage continuous emotion recognition framework is given in Figure 1. In dyadic interaction setup two participants, 1 and 2, interact, and we have speech and body motion recordings from these two participants. We extract frame level speech and body motion features and represent them as $f_i^S(k)$ and $f_i^M(k)$ at frame k for i-th participant. Correspondingly, $a_i^A(k)$, $a_i^V(k)$ and $a_i^D(k)$ are defined as the underlying activation, valence and dominance attributes for the i-th participant, respectively.

At the first stage, we obtain a window-level summarization of speech, motion and multimodal features for both participants independently to estimate their corresponding emotion attributes in a 3-class recognition setup. Subsequently, we exploit Gaussian Mixture Regression (GMR) to construct a statistical mapping between the underlying observed summarized speech, motion and multimodal features and the hidden window-level emotional attributes, a. After GMR, the estimated attribute, $\hat{a}$, is low-pass filtered for smoothing as in [11, 14].

In the second stage, we aim to improve the first stage estimates by performing CER of the i-th participant using her/his speech and body motion modalities as well as using, $\hat{a}_j$, the estimated affective attribute(s) of the j-th participant as a feature. This can improve CER because of the cross-emotional relationship between the speakers. GMR is used at the second stage again for the regression analysis. In essence, we validate the predictive ability of the our first stage cross-speaker AVD estimates by re-iterating them for CER to generate our second stage updated estimates, $\widehat{a\hat{a}}$. The use of cross-subject side emotional information for CER is motivated by observing the correlation between cross-subject affective attributes.

2.1. Database

In this work, we use the USC CreativeIT database, which is a multimodal database of theatrical improvisations [22, 23]. Each interaction on average has a length of 3.5 minutes and contains two recordings, one for each actor in a pair. The working database comprises of five sessions after aligning the three emotion attributes over a common time range with the audio and motion data, and removing noisy recording and bad annotations. Since our work is formulated for a cross-subject analysis, we require emotional attributes from both participants to evaluate the proposed system. Thus, after eliminating some recordings, which do not have emotional attributes for both participants, our final experimental dataset includes 65 time-synchronized and AVD aligned recordings. We use the mean inter-annotator ratings of three human annotators as the ground truth attributes. Also note that each session is performed by a unique pair of actors and therefore, session independence implies speaker independence.

2.2. Feature extraction and window-level summarization

Frame level acoustic features are extracted as 39 dimensional feature vectors including energy, the first 12 MFCCs together with the first and second time derivatives. Frame level motion feature vector is 24 dimensional and includes the Euler rotation angles in directions (x,y,z) of the arm and forearm joints together with their first derivatives over each frame. Acoustic features are computed every 16.67 ms with 8.33 ms overlap to match the motion capture rate.

In our window-level framework, we perform CER over windows of size 3 sec with an overlap of 1 sec between adjacent windows to generate the summarized speech, motion and multimodal feature vectors. We extract frame level feature vectors over the temporal duration of the window and construct matrices of features as $F_n^S = [f_1^S, ..., f_K^S]$ for the n-th window with dimensions $39 \times K$. Similarly, our motion feature matrix is constructed as $F_n^M = [f_1^M, ..., f_K^M]$ with dimensions $24 \times K$. Statistical functionals are computed over feature frames within the temporal duration of each window to provide a down-sampled statistical representation as in [11]. The resulting matrices of speech, motion and multimodal features are reduced by applying Principal Component Analysis (PCA) retaining 50 most informative principal components that corresponds to more than 90% of the total variance on average. We incorporate dynamic information by augmenting the PCA summarized features with their first order temporal derivatives. Speech, motion and multimodal data along with the emotional attributes are z-normalized using the global means and standard deviations of the dataset. Finally, we define our summarized speech, motion and multimodal feature vectors as h^S, h^M and h^{SM}, respectively. Note that in the proposed CSCER system, the cross-speaker emotion descriptor is appended to the feature set, which increases the feature dimensionality by one.

Since emotional attributes change slowly in comparison to audiovisual data, the resulting window-level covariance matrices may become singular causing the GMR to fail. One possible way to ensure increased signal variability while preserving signal spectral characteristics is to selectively add a small amount of noise to the slow-varying signal. Therefore, we add additive white Gaussian noise (AWGN) noise with a signal to noise ratio

Table 1: *Cross-speaker inter-attribute Pearson's correlation based on ground truth activation(A), valence(V) and dominance(D) attributes.*

		Participant 1		
		a_1^A	a_1^V	a_1^D
Participant 2	a_2^A	0.5783	-0.2887	0.1283
	a_2^V	-	0.2248	-0.1974
	a_2^D	-	-	-0.2617

(SNR) of 200dB in the SEI to increase the data variance.

We perform a pre-analysis to access the correlation between the emotional attributes of two participants engaged in a dyadic interaction. We analyze the correlation between mean inter-annotator agreements across sessions as ground truths for the three emotional attributes of USC CreativeIT dataset. As presented in Table 1, there is a significant correlation between the activation of the two speakers. The cross-attribute correlation generally seems insignificant for valence and dominance of both speakers since they are slightly correlated. Based on this analysis, we expect very slight improvement in estimation of valence and dominance attributes under our CSCER framework. Nonetheless, the inter-dependencies may still render slight improvements. On the other hand, ee expect activation prediction to reflect most significant advantage of utilizing cross-subject emotional dependencies as activation of both speakers seems promisingly correlated. In such an analysis, it should be kept in mind that the annotations are context dependent and subjective to dataset size thus it is hard to generalize inter-attribute correlations across datasets.

2.3. Cross-subject continuous emotion recognition (CSCER)

Continuous emotion recognition, which performs estimation of the affective AVD attributes, is formulated using the GMR framework as

$$\hat{a}(h) = \arg\max_a P(a \mid h, \lambda(h, a)), \quad (1)$$

where $\lambda(h, a)$ is the joint Gaussian mixture density of the observation and affect variables, h and a.

In the first stage, we adopt the GMR framework to estimate the affective AVD attributes using speech ($\hat{a}_i(h^S)$), motion ($\hat{a}_i(h^M)$) and multimodal ($\hat{a}_i(h^{SM})$) observations for the i-th participant.

In our second stage CSCER, we estimate attribute of the i-th participant using her/his speech/motion/multimodal data as well as using the estimated cross-subject affective attribute of the j-th participant. Note that the observation space dimension increases by one, and the second stage GMR estimation becomes,

$$\widehat{aa}_i(h_i, \hat{a}_j(h_j)) = \arg\max_a P(a \mid h_i, \hat{a}_j(h_j), \lambda(h, a)). \quad (2)$$

In an independent experimental setup, we also deploy ground truth from human agreements on AVD attributes as the true side information to set an upper performance bound for the second stage estimates. As the true side information is not available in a realistic scenario, it sets theoretically an upper performance benchmark. We can define the second stage CER estimate with true side information as $\widehat{aa}_i(h_i, a_j)$.

We perform a comparative analysis of CSCER estimates with the first stage (baseline) results and our ideal true side information based upper bound reference. For each of the estimated attributes, we deploy individual i.e. (a) activation, (b) valence and (c) dominance descriptions. In general, our experimental setup tests the predictive ability of intensity, pleasure and control of each participant on the prediction of emotional experience of other with and without supplementary cross-subject emotional descriptions.

3. Experimental evaluations

We devise three different experimental setups based on speech, motion and a multimodal data. The absolute relationship of features to emotional attributes is ensured thereby we report the absolute weighted mean correlation across sessions. It is empirically deduced that this approach is not sensitive to increased number of mixtures. We conduct five-fold-leave-one-session out cross-validation for testing.

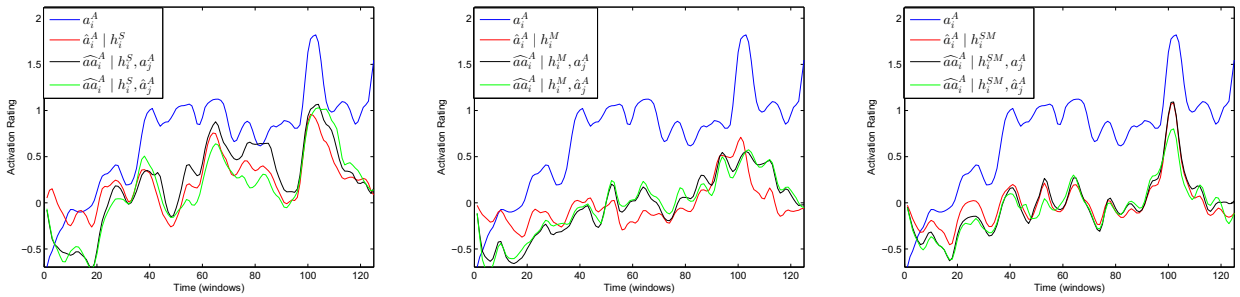
Table 2 presents CSCER results with true side information. As compared to the first stage, wherein AVD estimation is based on speech, motion and multimodal data only (bottom-most row), the true side information delivers considerable improvement for all the three estimated attributes. The advantage of cross-speaker emotional dependencies is best captured in activation as expected by our pre-analysis. While no significant dependency is observed between valence-dominance and dominance-activation, we observe that the slight correlation between activation-valence and vice versa shows advantage in activation and valence estimation over the first stage estimates. We obtain the highest improvements with like-attribute side AVD information for diagonals of Table 2, and this observation is consistent over all three modalities. We expected a slightly low performance in the multimodal case as reported due to overfitting because of low sample to feature ratio.

Table 3 presents CSCER results with estimated side information. We observe that our first stage estimate based CSCER renders performance that approaches the maximal bound validating a strong inter-subject emotional dependency that holds potential for improved CER. To validate our approach in a practical setting where true cross-subject emotions are not available on run-time, our goal was to obtain correlation close to that using the true side descriptions. We observe that our cross-subject estimated side information from the first stage approaches upper bounds set by true side information from Table 2. We obtain best results with speech followed by motion and multimodal data. Lower motion based performance may be due to arm and forearm features not reflecting sufficient movements in dyadic interactions. The correlation with estimated side information slightly exceeds the upper bound set by true side information. This is due to the unavoidable need of interpolation on the estimated data to align two partner sides and deliberate noise addition in the experimental process. Here again, use of cross-subject emotions is best captured in activation dimension.

Figure 2 presents a comparative performance analysis of activation recognition under CSCER using speech, motion and multimodal cues, conducted on a sample interaction. We obtain highest performance with speech and speech couplings as shown in Figure 2a. Additionally, we observe that the results achieved by the first stage are improved when we incorporate the cross-subject side information.

Table 2: *Cross-subject continuous emotion recognition (CSCER) at the window-level of AVD with ground truths as true side emotional information based on speech, motion and multimodal cues. We present mean absolute correlation values between estimated emotional curve and the ground truth.*

		Estimated Emotion								
		h^S			h^M			h^{SM}		
		$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$	$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$	$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$
Cross-subject side information	a_j^A	**0.6321**	0.3290	0.2526	**0.5113**	0.3135	0.2441	**0.5017**	0.2963	0.2504
	a_j^V	0.5164	0.3220	0.3530	0.3932	0.3559	0.3267	0.4734	0.3104	0.2609
	a_j^D	0.5106	0.3173	0.3675	0.3366	0.3446	0.3466	0.4681	0.2976	0.2870
	Stage1	0.5265	0.2896	0.3088	0.3686	0.2891	0.2350	0.5275	0.2913	0.2615



(a) *Speech* (b) *Motion* (c) *Multimodal*

Figure 2: *Comparison of (i) ground truth for Activation with estimation using (ii) modality data only, (c) modality data and true side information, and (d) modality data and estimated side information.*

Table 3: *CSCER at the window-level of activation, valence, dominance with estimated ground truths (Stage1) as side emotional information based on speech and motion cues. We present mean absolute correlation values between estimated emotional curve and the ground truth.*

	Estimated Emotion		
	$\widehat{aa}_i^A \mid \hat{a}_j^A$	$\widehat{aa}_i^V \mid \hat{a}_j^V$	$\widehat{aa}_i^D \mid \hat{a}_j^D$
h_i^S	**0.6189**	0.3441	0.3330
h_i^M	0.5079	0.3645	0.3484
h_i^{SM}	0.5361	0.3374	0.3240

4. Conclusions and future work

The proposed CSCER framework exploits cross-subject side information to improve CER performance in dyadic interaction setups. Our two-stage experimental setup initially performs emotion estimation using speech and motion data and then explores the effect of incorporating cross-attribute side information in a realistic cross-subject setting using speech, motion and multimodal data. In general, the benefit of cross-emotional dependency is best captured in the activation dimension. We find that the estimation of emotional attributes improves significantly with activation descriptions and marginally with valence and dominance. We validate our approach by demonstrating that the estimated attributes hold potential in achieving upper bound set by true side information. Our experimental setup offers flexibility to incorporate several other features such as synchrony for improved CER. As a future study, we aim to conduct the current analysis on our in-house JESTKOD database that is also structured to address affective dyadic interactions in agreement and disagreement scenarios [24].

5. References

[1] S. Tomkins, *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962, vol. 1.

[2] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[3] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[4] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[5] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[6] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[7] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2007, pp. 375–388.

[8] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[9] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[10] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2288–2291.

[11] 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.

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[13] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[14] H. Khaki and E. Erzin, "Use of agreement/disagreement classification in dyadic interactions for continuous emotion recognition," *Proceedings of Interspeech, San Francisco, USA*, 2016.

[15] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[16] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, pp. 5–32, 2003.

[17] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[18] E. Bozkurt, S. Asta, S. Özkul, Y. Yemez, and E. Erzin, "Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3652–3656.

[19] 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, 2013.

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

[21] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[22] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[23] 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.

[24] E. Bozkurt, H. Khaki, S. Keçeci, B. B. Türker, Y. Yemez, and E. Erzin, "The jestkod database: an affective multimodal database of dyadic interactions," *Language Resources and Evaluation*, 2016.