Lang Resources & Evaluation (2017) 51:857–872
DOI 10.1007/s10579-016-9377-0

ORIGINAL PAPER

The JESTKOD database: an affective multimodal database of dyadic interactions

Elif Bozkurt[1] · Hossein Khaki[1] · Sinan Keçeci[1] · B. Berker Türker[1] · Yücel Yemez[1] · Engin Erzin[1]

Published online: 28 November 2016

Abstract In human-to-human communication, gesture and speech co-exist in time with a tight synchrony, and gestures are often utilized to complement or to emphasize speech. In human–computer interaction systems, natural, affective and believable use of gestures would be a valuable key component in adopting and emphasizing human-centered aspects. However, natural and affective multimodal data, for studying computational models of gesture and speech, is limited. In this study, we introduce the JESTKOD database, which consists of speech and full-body motion capture data recordings in dyadic interaction setting under agreement and disagreement scenarios. Participants of the dyadic interactions are native Turkish speakers and recordings of each participant are rated in dimensional affect space. We present our multimodal data collection and annotation process, as well as our preliminary experimental studies on agreement/disagreement classification of dyadic interactions using body gesture and speech data. The JESTKOD database provides a valuable asset to investigate gesture and speech towards designing more natural and affective human–computer interaction systems.

Keywords Gesture · Speech · Affective state tracking · Human–computer interaction · Dyadic interaction

✉ Engin Erzin
eerzin@ku.edu.tr

[1] Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Istanbul, Turkey

1 Introduction

Social signals are perceivable stimuli that, either directly or indirectly, convey information concerning social actions, interactions, attitudes, emotions and relations (Vinciarelli et al. 2012). 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 of Carletta (2007), McCowan et al. (2005), Hillard et al. (2003) and Galley et al. (2004), political debates of Kim et al. (2012), Bousmalis et al. (2011) and Vinciarelli et al. (2009), theatrical improvisations of Metallinou et al. (2015) and broadcast conversations of Grimm et al. (2008) and Wang et al. (2011).

Large collections of interaction data, recorded in naturalistic settings, are needed to develop and evaluate statistical models of multi-party conversations. In this paper, we present a literature survey on multimodal databases of dyadic interactions and introduce our JESTKOD database, which includes multimodal affective recordings of spontaneous dyadic interactions under agreement and disagreement scenarios. The JESTKOD database contains high-quality audio, video and motion-capture recordings of dyadic interactions and provides a valuable asset to investigate gesture and speech signals in natural and affective interaction settings. It has potential to create useful models for affective human–computer interaction systems. We also investigate speech and body motion modalities to model agreement and disagreement in dyadic interactions, and present some early classification results.

Poggi et al. (2010) define the notion of agreement as a relation of identity, similarity or congruence between the opinions of two or more people and show that agreement can be communicated by different modalities as speech and body signals. Bousmalis et al. (2009) summarize cues for agreement and disagreement. Facial expressions, head gestures, gaze, laughter, and body posture are among the most preferred cues used for the analysis purposes. Most of the techniques available can only deal with a very limited number of hand gestures, e.g. hand cross, forefinger raise. Furthermore, most of the existing databases require locating the hand, tracking it and then interpreting cues for agreement or disagreement. On the other hand, the multimodal JESTKOD database provides joint angle rotation and position information for full body and for both participants of a dyadic interaction.

Manually annotated hand and head gestures together with speech prosody are used for agreement/disagreement classification on a political debate dataset by Bousmalis et al. (2011), where support vector machines (SVM), hidden Markov models (HMM), and hidden conditional random fields (HCRF) were employed as classifiers. The HCRF classifier with multimodal data achieves 64.22% accuracy rate for the agreement/disagreement classification. Kim et al. (2012) investigate an extreme case of disagreement (conflict) using prosodic and conversational cues on a political debate dataset, as well. They report performances of an SVM classifier with recall rate up to 71.9% for low, medium, and high level conflict classes.

Dyadic interaction requires social interactions, such as coordination and calibration. Bavelas et al. (1995) point out that person who has the speaking turn

in a dialog constantly includes the addressee, and hand gestures can help the interlocutors coordinate their dialog and serve the special conversational demands of talking in dialog rather than in monologue. Supporting this point of view, Yang et al. (2014) show that individuals' attitude as well as the interaction type can be predicted as friendly or conflictive using only the dynamics of the hand gesture patterns over an interaction. Moreover, humans are able to distinguish statements of agreement from disagreement and neutral utterances on the basis of low-level nonverbal cues alone (Mehu and Maaten 2014).

Emotions are a part of everyday communication, and only few works, including (Yang et al. 2014; Yang and Narayanan 2014, 2016), have attempted to analyze gestures in affective human–human interactions. This is partly so because relevant datasets, which would help to identify discriminative combinations of multimodal cues, are scarce, and because there is a gap of relevant annotated data that can be used for such analyses. JESTKOD is a multimodal speech and body motion database, which is collected in natural dyadic interaction settings. Dyadic interactions are set with agreement and disagreement tasks to create affective variability. Recordings of each participant in the JESTKOD are annotated using the dimensional attributes of activation, valence and dominance (AVD). The main contribution of the JESTKOD is to present speech and motion data in natural dyadic interactions. Furthermore, use of the agreement/disagreement task to create the targeted variability in AVD attribute space is a secondary contribution. In this sense, the JESTKOD database provides a valuable asset for investigating gesture and speech signals in affective interaction scenarios. As well it constructs a rich body motion repertoire in natural dyadic interactions that can contribute to speech driven gesture synthesis and animation.

This paper is organized as follows. In Sect. 2 we present a literature review on expressive interactions from computational aspects of speech and gestures. Then, we describe our multimodal data collection process and extent of the JESTKOD database. In Sect. 3 we present early experimental studies on agreement/disagreement classification of dyadic interactions using body gesture and speech data. Finally, in Sect. 4 we provide conclusions.

2 Multimodal dyadic interaction database

2.1 Literature review

There are a variety of multimodal databases that contain continuous affect annotations, which are publicly available for research purposes. The HUMAINE database includes a large collection of multimodal naturalistic and induced recordings (Douglas-Cowie et al. 2007). The SEMAINE database consists of audio–visual data in the form of conversations between participants and a number of virtual characters with particular personalities (McKeown et al. 2012). The acted audio–visual MSP-IMPROV database investigates emotional behaviors during conversational dyadic improvisations (Busso et al. 2016). We note that, although motion capture technologies are becoming widely available, there exist only a

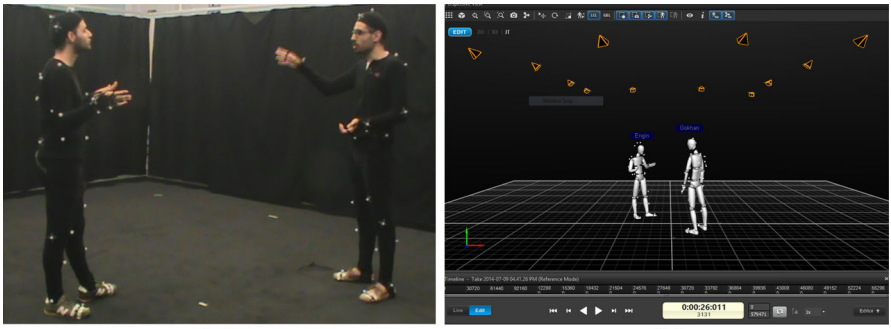

Fig. 1 Samples from the video and motion-capture recordings of the JESTKOD database. **a** Video recordings and **b** motion-capture

limited number of audio–visual databases which also include 3D motion data for modeling bodily gestures. Heloir et al. (2010) explore technical setups, scenarios and challenges in building a motion capture database for virtual human animation. Busso et al. (2008) present their interactive emotional dyadic motion capture (the USC IEMOCAP) database, 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 (Metallinou et al. 2010, 2015). The USC CreativeIT database is annotated using the valence, activation and dominance attributes, as well as the theater performance ratings such as interest and naturalness. Since interaction performances of the USC CreativeIT database are theatrical, speech and body gestures are rather amplified and exaggerated in their settings.

2.2 JESTKOD database

Our main motivation to construct the JESTKOD database is to collect multimodal speech and body gesture data in natural and affective dyadic interactions. It can provide a valuable asset to investigate use of gestures during spoken interactions. The JESTKOD database consists of dyadic interaction recordings of 10 participants, 4 females and 6 males, ages from 20 to 25. For data collection, each participant was paired with the same interlocutor for both agreement and disagreement scenarios. Hence 10 participants were grouped in 5 pairings. Recordings for each pairing were collected in 5 sessions, all in Turkish. In each session, there are 19–23 clip recordings of 2–4 min, where in each clip participants discuss on a topic that they agree or disagree in dyadic interaction. The total duration of the recordings is 259 min. Recordings were captured by a high-definition video camera, a full body 3D motion capture system and Sony condenser tie-pin microphones. We used the *OptiTrack Flex 13*[1] system and the *Motive*[2] software for full body motion capture,

[1] Flex 13 system—http://www.optitrack.com/products/flex-13/.

[2] Motive—optical motion capture software http://www.optitrack.com/products/motive/.

Table 1 Session-level summary of topics in the JESTKOD database

Session	Agreement scenario	Disagreement scenario
1	Movies, world cuisine, holiday resorts, TV series	Soccer, mathematics, game consoles, PC games
2	Soccer, world cuisine, movies, literature	Geography, holiday resorts, theater, dance
3	Movies, sports, PC games, music, world cuisine	Movies, history, animals, education
4	World cuisine, holiday resorts, science-fiction, history, theater, cities	Soccer, movies, PC games, TV series, literature, physics
5	Movies, languages, PC games, cities, game consoles	Movies, sports, holiday resorts, nutrition, musicals

which consists of 12 infrared cameras capturing 21 body joints at 120 fps with a resolution of 1280 × 1024. A sample scene from the video recordings and a screen shot from the motion capture software are given in Fig. 1.

Topics of the dyadic interactions were set by the moderator of the session using a preliminary information form, which was filled by all participants before the recordings. In the preliminary information form, participants were asked to state their favorite and disliked soccer teams, food, restaurants, world cuisines, computer games, movies, operating systems, game consoles, etc. 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. The moderator instructed participants to engage in the conversation in their natural daily manners. Distributions of the topics in each session are summarized in Table 1. Note that, the same topic may appear in both agreement and disagreement interactions. For example, the soccer topic can initiate a strong disagreement interaction between participants supporting different teams. On the other hand, participants supporting the same team can engage in a congruent conversation. In the JESTKOD database we have 56 dyadic interactions in agreement and 42 dyadic interactions in disagreement with total durations of 154 and 105 min, respectively.

2.2.1 Continuous annotation of recordings

In the psychology literature, three approaches have been introduced for modelling affect: categorical, dimensional and appraisal-based approaches (Grandjean et al. 2008). The categorical approach mostly assumes a small number of universal emotion classes such as, anger, disgust, fear, happiness, sadness, and surprise (Ekman and Friesen 1975). However, this approach is limited since, humans have the ability to perceive and express more complicated affective states like depression (Russell 1980). On the other hand, dimensional approaches represent affect on a continous scale, which is useful for explaining non-categorical, complex affective states and also affective state transitions during human interactions. Dimensional approaches mostly concentrate on representation of affect in the two dimensional

space as in activation and valence domains. Additionally, some researchers also include the third dimension as the dominance domain into their analyses. In these models, affect is described along the active–passive, positive–negative and dominant–submissive dimensions (Mehrabian 1996). A detailed overview on the continuous representation of affect can be found in Gunes et al. (2011). The third approach claims that emotions are elicited by continuous subjective evaluations (appraisals) of the outside world (Scherer et al. 2001). However, it is still an open research area how to use the appraisal based affect approach for automatic measurement of affect, since it is challenging to evaluate such a complicated, multi-partite signal.

In this paper, the JESTKOD database is constructed to represent natural affective variability in the dyadic interactions. For data collection, participants were asked to engage in a conversation on a given topic, which triggered agreement or disagreement in an open-ended and continuous interaction. Since categorical emotion classes were not in the target, dimensional attributes of activation, valence and dominance (AVD) were used to assess affective variability of dyadic interactions. Annotators rated each participant in the recordings using dimensional attributes of AVD in [0, 1] range. Ratings were collected using the general trace program (*GTrace*) from Queen's University (Cowie et al. 2011). Annotation effort was carried out separately for each participant in the recordings and for each dimensional attribute. A joystick interface was used with the *GTrace* software to deliver continuous-time ratings of the activation, valence and dominance attributes. A total of six annotators contributed to collect three sets of ratings for valence and dominance, and four sets of ratings for activation attribute. The annotators were selected among undergraduate engineering students who took a pre-training program on the annotation task and performed well in this task. In the pre-training program candidates were instructed on definitions and characteristics of the attributes, and they practiced annotation task on three sample videos for each attribute. They received feedback on their annotation performance and repeated the task several times. The average duration of the pre-training program was around one hour. Note also that total duration of the database is 259 min. Considering a real-time rating factor of 5 and ratings of two participants for each recording, the total duration of the rating effort is 2590 min for each attribute. Total amount of ratings are maximized within the funding limits of the supporting research project.[3]

A consensus rating, which can be defined as the summary of all ratings over the annotators, is extracted using correlation as a basis of concordance between annotators as in Metallinou et al. (2015). Each recording of a participant can then be represented with the consensus rating in AVD attributes. The consensus rating is extracted over temporal windows of duration 15 s, which are temporally 50% overlapping. The window-level ratings are linearly combined over the overlapping segments to set the final consensus rating for each recording of a participant. Linear combination over overlapping segments performs a smoothing over the consensus ratings to eliminate sharp transitions at the window boundaries. Note that in combining ratings of the annotators, some ratings may appear inconsistent with the

[3] The JESTKOD project is supported by TÜBİTAK under Grant Number 113E102.

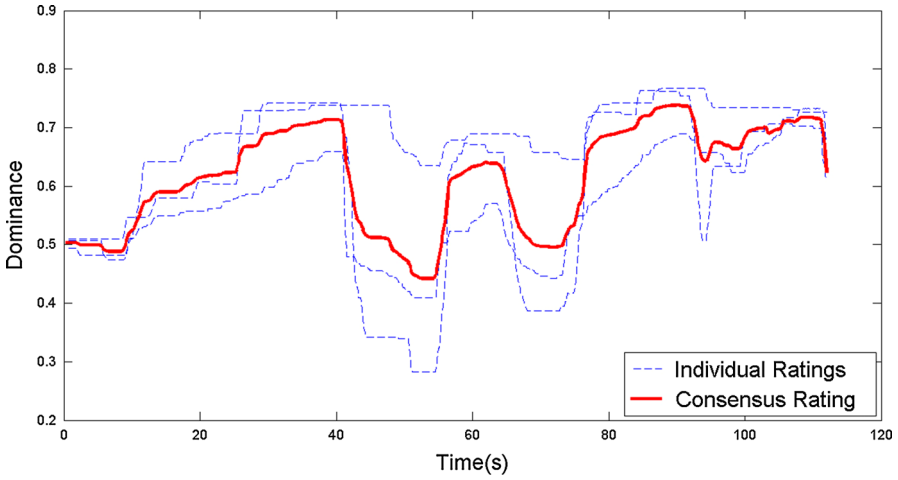

Fig. 2 Sample ratings of annotators and the extracted consensus rating

rest. This is common issue in emotional labeling with categorical labels, where a consensus is often computed based on majority voting and minority labels are ignored. Metallinou et al. (2015) extend majority voting concept to combine continuous emotional ratings by setting a cut-off correlation threshold for an acceptable consensus. In this study, we adopt a similar approach and empirically set the threshold to 0.4. Window-level consensus rating computation is carried out to prune ratings of annotators in low agreement with the resulting consensus rating. For pruning of unreliable ratings, correlation coefficients between the consensus rating and individual ratings of the annotators are calculated for each window. If all the correlation coefficients are higher than the predefined threshold of 0.4, the consensus rating is fixed. Otherwise, the least correlated rating is dropped and the consensus rating extraction is repeated for the window. This pruning process is carried out while the consensus rating computations are using more than two annotators. In case ratings with low agreement are pruned to have only ratings of two annotators, the consensus rating is fixed over the ratings of the last two annotators. Figure 2 presents a sample plot with three individual ratings and the extracted consensus rating.

2.2.2 Statistical analysis of annotations

In order to assess the quality of the consensus rating obtained by the procedure described in Sect. 2.2.1, the Pearson's correlation coefficients are calculated between the consensus rating and individual ratings of the annotators over the temporal windows. The mean and SD values of correlation coefficients are reported in Table 2 for each of the AVD attributes. Note that the mean values of the correlation are all positive, which is a good indication of agreement between individual ratings and the consensus ratings.

Table 2 Mean and SD of the Pearson's correlation between the consensus rating and individual ratings for the activation, valence and dominance attributes under different interaction types

	Activation		Valence		Dominance	
	Mean	SD	Mean	SD	Mean	SD
Agreement	0.546	0.279	0.598	0.275	0.756	0.230
Disagreement	0.568	0.265	0.530	0.315	0.718	0.258
All	0.557	0.270	0.564	0.300	0.737	0.240

Table 3 Ratio of windows (in percent) having correlations higher than 0.4 between the consensus rating and individual ratings

Activation	Valence	Dominance
76	66	82

The ratio of windows with high correlation values, i.e. higher than the threshold value 0.4, is also considered as a statistical metric to evaluate the consensus rating. In Table 3, the ratio of windows with high correlation values are reported. Note that dominance has the highest level of concordance with 82%. This means that the consensus rating is correlated to all individual ratings with a correlation value higher than 0.4 for 82% of the windows. Equivalently, for 18% of the windows the consensus rating is computed over two annotators and at least one of the correlation coeffients between the consensus and individual ratings is less than 0.4 for the dominance attribute. Activation has the second highest level of concordance with 76%, and finally valence has the lowest level of concordance with 66%. Even with the valence attribute, the annotators are in strong concordance with each other on two thirds of the database.

Another interesting investigation is to observe distributions of the consensus AVD ratings for agreement and disagreement interaction types. AVD histograms over the consensus ratings are computed for each attribute separately for agreement and disagreement. Figure 3 plots the computed histograms. Activation and dominance do not convey significant distribution differences between agreement and disagreement. However, histograms of the consensus valence ratings differ significantly under agreement and disagreement interactions. Note that, as expected, the histogram in the case of agreement is shifted along the positive valence axis when compared to disagreement. Similarly, Fig. 4 plots the joint histograms of the consensus AVD rating pairs. Note that activation–valence and dominance–valence consensus rating distributions significantly differ with the interaction type.

We utilized the Kullback–Leibler divergence (KLD) to quantify statistical difference between agreement and disagreement distributions. The KLD and symmetric KLD can be defined respectively as,

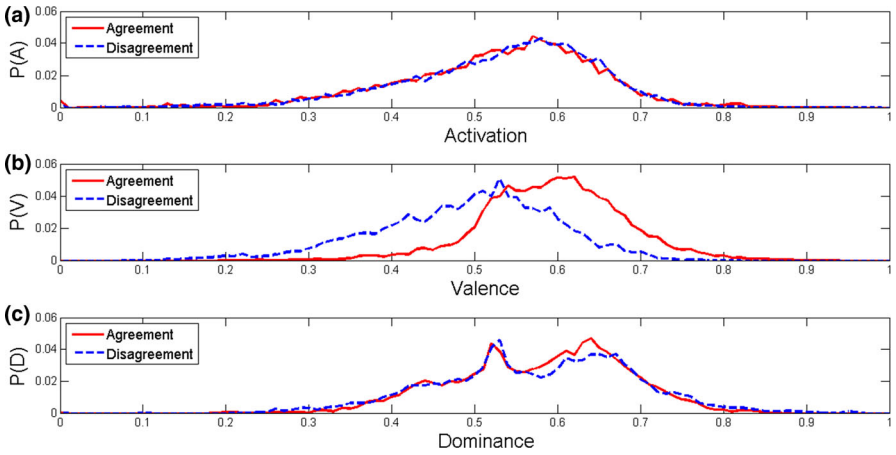

Fig. 3 Histograms of the consensus ratings for activation, valence and dominance attributes under agreement and disagreement interactions

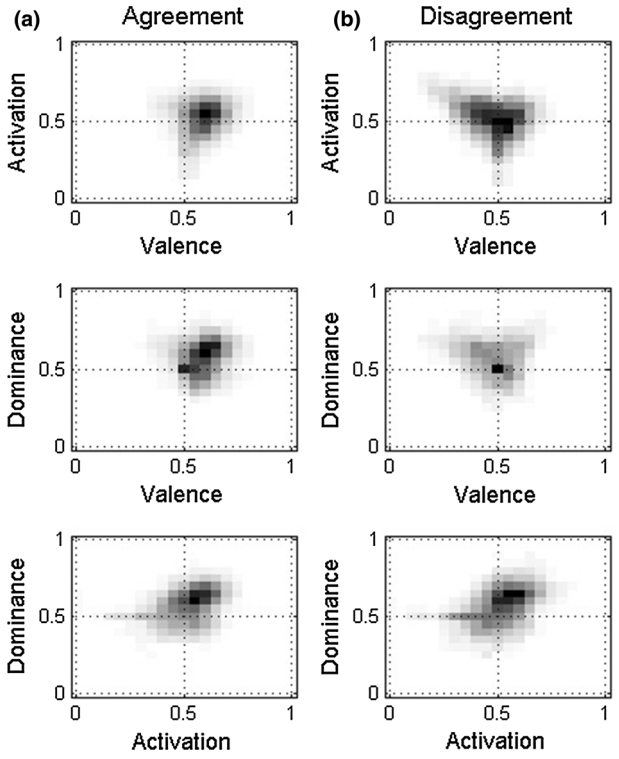

Fig. 4 Joint histograms of the consensus rating for **a** agreement and **b** disagreement interactions over activation–valence, dominance–valence and dominance–activation attribute spaces

Table 4 Symmetric KLD distances between agreement/disagreement interactions for the consensus activation, valence and dominance ratings: diagonals are over the marginal distributions and off-diagonals are over the joint distributions

	$D_{KL}(P_A, P_D)$		
	Activation	Valence	Dominance
A	0.033	3.042	1.157
V	3.042	1.184	4.697
D	1.157	4.697	0.208

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

and

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

where $P_A()$ and $P_D()$ are probability distributions respectively over agreement and disagreement scenarios, and k runs over the sample space of consensus ratings. Table 4 presents on the diagonal the symmetric KLD distances between the distributions over agreement and disagreement scenarios for activation, valence and dominance, whereas the off-diagonal values represent the distance between joint distributions. All KLD distances are computed using the same number of bins for each attribute, hence they are comparable with each other in magnitude. The agreement and disagreement distributions have the strongest difference for the valence attribute with a KLD distance of 1.184. The KLD distance for the activation attribute is 0.033, which does not indicate significant statistical difference between agreement and disagreement, whereas dominance presents a KLD distance of 0.208. As for the joint distributions, dominance–valence yields the highest KLD distance as 4.697 while the activation–dominance distribution attains a distance of 1.157, which is lower than the KLD distance over the marginal valence distribution.

3 Agreement/disagreement classification

An important aspect of the JESTKOD database is the multimodal speech and body motion capture under agreement/disagreement scenarios. We pose a two-class dyadic interaction type (DIT) estimation problem of agreement and disagreement classes from speech and motion modalities to assess their relationship. A block diagram of the classification system is given in Fig. 5. Speech and motion streams of the two participants of the dyadic interaction are the inputs to the feature extraction block. Then frame level features of speech and motion are fed into utterance extraction block to compose temporal collection of feature vectors. Joint

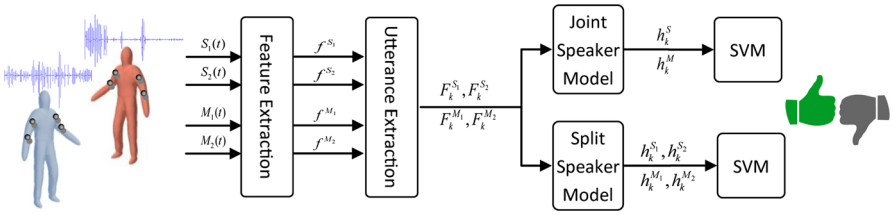

Fig. 5 Block diagram of the agreement/disagreement classification system

and split speaker models perform a statistical summarization on the utterance-level features. Finally, SVM classifiers perform DIT estimation over the summarized feature representations. We describe these blocks in the following subsections in detail.

3.1 Feature and utterance extraction

The speech signal of each participant is processed over 20 ms windows with 10 ms frame shifts to extract 13 dimensional MFCC feature vector together with its first and second order derivatives, f^S. On the other hand frame level motion feature vector f^M for each participant is extracted from the Euler rotation angles in directions (x, y, z) of the arm and forearm joints together with their first derivatives. Note that frame rates for the speech and motion modalities are respectively 100 and 120 fps.

In the utterance extraction frame level feature vectors are collected over the temporal duration of the utterance. The silence frames are filtered out for the speech modality. The speech feature matrix is constructed as $F_k^S = [f_1^S \ldots f_{N_S}^S]$ for the kth utterance with dimensions $39 \times N_S$. Similarly the motion feature matrix is then constructed for each participant as $F_k^M = [f_1^M \ldots f_{N_M}^M]$ for the kth utterance with dimensions $24 \times N_M$ without silence filtering.

3.2 Feature summarization

A statistical summarization is performed to map the high dimensional utterance level feature matrices to low dimensional feature representations. We use statistical functionals mean, SD, 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 as defined in Metallinou et al. (2013).

Using the statistical functionals, two separate models of feature representation are constructed, namely joint and split speaker models. In the joint speaker model (JSM), the features of two participants in each session are collected together to apply statistical summarization. Hence for the speech modality, the combined feature matrices of the kth utterance, $[F_k^{S_1} F_k^{S_2}]$, are fed into statistical summarization to extract summarized feature representation h_k^S, where $F_k^{S_1}$ and $F_k^{S_2}$ denote speech feature matrices of the first and the second participants in the session. Similarly,

summarized feature representation h_k^M for the motion modality is extracted from $[F_k^{M_1} F_k^{M_2}]$ using statistical summarization, where $F_k^{M_1}$ and $F_k^{M_2}$ denote motion feature matrices of the participants in the session.

In the split speaker model (SSM), statistical summarization is applied for each participant in a given session and then the summarized features are combined to represent speech and motion modalities. Hence the summarized feature representations $h_k^{S_i}$ and $h_k^{M_i}$ are extracted for the speech and motion modalities respectively from $F_k^{S_i}$ and $F_k^{M_i}$, $i = 1, 2$. Then the combination of $h_k^{S_1}$ and $h_k^{S_2}$ is used as the summarized feature representation of the speech modality for the given session. Similarly, the combination of $h_k^{M_1}$ and $h_k^{M_2}$ is used as the summarized feature representation of the motion modality.

3.3 SVM classification

Support vector machine (SVM) is used as a binary classifier for the DIT estimation. Let $C(h)$ denote the SVM classifier, which is using the feature vector h. Then the unimodal and multimodal classifiers using the JSM can be defined as $C(h^S)$, $C(h^M)$ and $C(h^S, h^M)$ respectively for speech, motion and joint speech-motion. Similar unimodal and multimodal classifiers using the SSM can be defined as $C(h^{S_1}, h^{S_2})$, $C(h^{M_1}, h^{M_2})$ and $C(h^{S_1}, h^{S_2}, h^{S_1}, h^{S_2})$.

3.4 Experimental evaluations

Leave-one-clip-out training is used within the JESTKOD database, where test is performed on one recording at a time and model training is performed on the remaining recordings. The estimation performance is calculated as the average of agreement and disagreement classification accuracies. Classification evaluations are performed at clip level and at utterance level. In the following two subsections clip level and utterance level classifications are discussed separately.

3.4.1 Clip level classification

In clip level classification, features of a clip are concatenated and summarized, and the DIT per clip is estimated. The classification accuracy results of the JSM and SSM experiments with unimodal and multimodal classifiers are presented in Table 5. In these two experiments, the classification results of the motion modality have lower accuracy, yet they contribute to the performances of the multimodal classifiers. Note that multimodal classifiers attain the highest classification rates. Furthermore comparing the results of JSM and SSM models, the split speaker model (SSM) does better for each modality compared to the joint speaker model (JSM).

3.4.2 Utterance level classification

The effect of the utterance duration on the classification accuracy is also investigated for DIT estimation. In the utterance level classification, the DIT

Table 5 Unimodal and multimodal classification accuracies for clip level DIT estimation

Method	Classification accuracy (%)
JSM	
$C(h^M)$	82.79
$C(h^S)$	83.61
$C(h^S, h^M)$	86.07
SSM	
$C(h^{M_1}, h^{M_2})$	79.51
$C(h^{S_1}, h^{S_2})$	89.34
$C(h^{S_1}, h^{S_2}, h^{M_1}, h^{M_2})$	90.16

estimation has been performed over overlapping utterances. The SSM classifiers having the highest clip level results are used over varying utterance durations. Here the duration is the total time of dyadic interaction, including silent and speech segments.

The classification accuracy of the utterance level experiments as a function of utterance durations in the range [5, 100] s with step of 5 s are depicted in Fig. 6. Note that when utterance duration is larger than 15 s, the multimodal accuracy is higher than 80% for the binary agreement and disagreement classification task. For the SSM classifiers when the duration is greater than 75 s, accuracy reaches to the clip level accuracy, which is around 90%. Furthermore the multimodal performance curve always has the highest accuracy.

3.5 Discussion

As observed in Table 5, the SSM outperforms the JSM with speech-only and multimodal classifiers but not with the motion-only classifier. In SSM, participants are modeled separately, however, in JSM, features of the two participants are combined and the overall statistics is calculated. The joint speaker model has the potential of averaging statistics of two different participants, and due to this mixing it may fall short to represent participants' characteristics, especially in speech modality. On the other hand, the better performance of the motion-only classifier under JSM suggests that the motion modality is not as personalized as the speech modality.

While Fig. 6 shows an overall increasing trend in the accuracy, the individual trends are not completely monotonic. This may be due to the variable duration silence segments within the utterances. Hence durations of the active speech can effect the quality of statistical features. Since time synchrony between speech and motion modalities is needed for the multimodal classification system, the multimodal system can not be tested over only active speech segments.

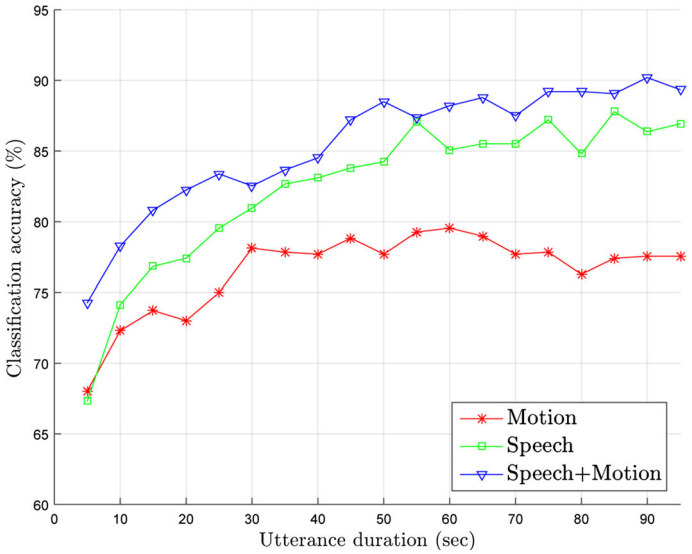

Fig. 6 Average agreement and disagreement classification accuracy as a function of utterance duration with SSM and statistical functionals

4 Conclusions and future work

In this study we have introduced the multimodal JESTKOD database comprised of speech, motion capture data and video recordings of continuously annotated affective dyadic interactions under agreement and disagreement scenarios. We have presented a detailed description of the data collection process, the annotation of recordings in the continuous affect domains as well as some preliminary results on agreement/dis-agreement classification of interactions over low level speech and motion capture representations. The JESTKOD database is composed of 5 sessions, where in each session a different dyad out of 10 participants engages in conversation on a topic selected by considering participants' shared or contradicting opinions. The total number of clips for agreement and disagreement scenarios is 56 and 42, respectively. In the JESTKOD database, each participant in each clip is annotated using dimensional attributes of activation, valence and dominance by a total of six annotators. The annotators' concordence is assessed to be fairly high by using a correlation based method. Furthermore, the dyadic interaction type (DIT) is estimated as agreement and disagreement by utilizing SVM classifiers based on statistical functionals that summarize temporal features. Our findings indicate that the low level speech features carry more discriminative clues than the motion features for DIT classification, whereas the multimodal features increase the accuracy of classification up to 90.16%. In addition, 15 s of utterance analysis window is enough to obtain an accuracy higher than 80% for the binary agreement and disagreement classification task when multimodal features are used.

The JESTKOD database provides a valuable asset to investigate gesture and speech signals in affective scenarios. In a recent study, Khaki and Erzin (2016) investigate the relationship between the continuous affect attributes activation, dominance and valence, and dyadic interaction type (DIT), and report improvements in estimating the valence attribute from speech and motion when DIT is available. As well, the JESTKOD provides a rich body motion repertoire in natural dyadic interactions that can contribute to speech driven gesture synthesis and animation. We finally note that samples from the JESTKOD database are available online.[4]

Acknowledgements This work is supported by TÜBİTAK under Grant Number 113E102.

References

Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, *21*(4), 394–405.

Bousmalis, K., Mehu, M., & Pantic, M. (2009) Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *3rd International conference on affective computing and intelligent interaction and workshops* (pp. 1–9).

Bousmalis, K., Morency, L., & Pantic, M. (2011). Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *IEEE international conference on automatic face gesture recognition and workshops (FG 2011)* (pp. 746–752).

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359.

Busso, C., Parthasarathy, S., Burmania, A., Abdel-Wahab, M., Sadoughi, N., & Provost, E. M. (2016). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, *10*(99), 1–1.

Carletta, J. (2007). Unleashing the killer corpus: Experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, *41*(2), 181–190.

Cowie, R., Cox, C., Martin, J. C., Batliner, A., Heylen, D., & Karpouzis, K. (2011). Issues in data labelling. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-oriented systems: The Humaine handbook* (pp. 215–244). Berlin: Springer.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., et al. (2007). The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In A. Paiva, R. Prada, & R. Picard (Eds.), *Affective computing and intelligent interaction. Lecture notes in computer science* (Vol. 4738, pp. 488–500). Berlin: Springer.

Ekman, P., & Friesen, W. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues. Spectrum books*. Englewood Cliffs: Prentice-Hall.

Galley, M., McKeown, K., Hirschberg, J., & Shriberg, E. (2004) 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* (p. 669). Association for Computational Linguistics, Stroudsburg, PA, USA.

Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition*, *17*(2), 484–495.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The vera am mittag German audio–visual emotional speech database. In *IEEE international conference on multimedia and expo* (pp. 865–868).

Gunes, H., Schuller, B., Pantic, M., & Cowie. R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE international conference on automatic face & gesture recognition and workshops (FG 2011)* (pp. 827–834). IEEE.

[4] The JESTKOD database—http://mvgl.ku.edu.tr/databases/.

Heloir, A., Neff, M., & Kipp, M. (2010). Exploiting motion capture for virtual human animation: Data collection and annotation visualization. In *Proceedings of the workshop on multimodal corpora: Advances in capturing, coding and analyzing multimodality* (pp. 59–62).

Hillard, D., Ostendorf, M., & Shriberg, E. (2003). 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 proceedings of HLT-NAACL 2003: Short Papers* (Vol. 2, pp. 34–36). Association for Computational Linguistics, Stroudsburg, PA, USA.

Khaki, H., & Erzin, E. (2016). Use of agreement/disagreement classification in dyadic interactions for continuous emotion recognition. In *Proceedings of Interspeech*, San Francisco, USA.

Kim, S., Valente, F., & Vinciarelli, A. (2012). Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5089–5092).

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 305–317.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261–292.

Mehu, M., & Maaten, L. (2014). Multimodal integration of dynamic audio–visual cues in the communication of agreement and disagreement. *Journal of Nonverbal Behavior*, *38*(4), 569–597.

Metallinou, A., Lee, C. C., Busso, C., Carnicke, S., & Narayanan, S. S. (2010). The USC CreativeIT database : A multimodal database of theatrical improvisation. In *Multimodal corpora: Advances in capturing, coding and analyzing multimodality (MMC)*.

Metallinou, A., Yang, Z., Lee, C. C., Busso, C., Carnicke, S., & Narayanan, S. S. (2015). The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, *50*(3), 497–521.

Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, *31*(2), 137–152.

Poggi, I., D'Errico, F., & Vincze, L. (2010). Agreement and its multimodal communication in debates: A qualitative analysis. *Cognitive Computation*, *3*(3), 466–479.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford: Oxford University Press.

Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the international conference on affective computing and intelligent interaction, ACII '09* (pp. 1–4).

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., et al. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), 69–87.

Wang, W., Yaman, S., Precoda, K., & Richey, C. (2011). Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 5556–5559).

Yang, Z., & Narayanan, S. S. (2014). Analysis of emotional effect on speech-body gesture interplay. In *Proceedings of Interspeech* (pp. 1934–1938). Singapore.

Yang, Z., & Narayanan, S. S. (2016). Modeling dynamics of expressive body gestures in dyadic interactions. *IEEE Transactions on Affective Computing*. doi:10.1109/TAFFC.2016.2542812.

Yang, Z., Metallinou, A., Erzin, E., & Narayanan, S. S. (2014). Analysis of interaction attitudes using data-driven hand gesture phrases. In *Proceedings of IEEE international conference on audio, speech and signal processing (ICASSP)* (pp. 699–703). Florence, Italy.