AGREEMENT AND DISAGREEMENT CLASSIFICATION OF DYADIC INTERACTIONS USING VOCAL AND GESTURAL CUES

Hossein Khaki, Elif Bozkurt, Engin Erzin

Multimedia, Vision and Graphics Lab,
Koç University, Istanbul, Turkey

hkhaki13,ebozkurt,eerzin@ku.edu.tr

ABSTRACT

In human-to-human communication gesture and speech co-exist in time with a tight synchrony, where we tend to use gestures to complement or to emphasize speech. In this study, we investigate roles of vocal and gestural cues to identify a dyadic interaction as agreement and disagreement. 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. Spectral features of vocal channel and upper body joint angles of gestural channel are employed to extract unimodal and multimodal classification performances. Both of the modalities attain classification rates significantly above the chance level and the multimodal classifier performed more than 80% classification rate over 15 second utterances using statistical features of speech and motion.

Index Terms— Gesticulation, speech, affective state tracking, human-computer interaction, Dyadic interaction

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, 3, 4], political debates [5, 6] and broadcast conversation [7].

Large collections of interaction data, recorded in naturalistic settings, are needed to develop and evaluate statistical models of agreement and disagreement in interactions. In this paper, we present a literature survey on multimodal databases of dyadic interactions and introduce our in-house JESTKOD database, which includes multimodal affective recordings of spontaneous dyadic interactions under agreement and disagreement scenarios. Then we investigate speech and body motion modalities to model agreement and disagreement in dyadic interactions. 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 for natural and affective human-computer interaction systems.

Bousmalis et al. summarize cues for agreement and disagreement in [8]. Facial expressions, head gestures, gaze, laughter, and body posture are among the most preferred cues used for the analysis purposes. However, spontaneous hand and body related cues have not been explicitly collected or modeled previously for agreement and disagreement analysis to the best of our knowledge. 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 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 in [6]. SVMs, hidden Markov models (HMMs), and hidden conditional random fields (HCRFs) are employed as the classifier. HCRF classifier with multimodal data achieves 64.22% accuracy rate for the agreement/disagreement classification. Kim et al. 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. 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 [9]. Supporting this point of view, Yang et al. show that individuals attitude as well as the interaction type as friendly or conflictive can be predicted using only the dynamics of the hand gesture phrases over an interaction [10].

We employ both unimodal and multimodal speech and arm motion features to classify agreement and disagreement in dyadic interactions using the JESTKOD database. We parameterize speech with mel-frequency cepstral coefficient (MFCC) and arm motion with Euler angle rotations. Then we use statistical functionals and i-vector representation for the summarization of the speech and motion features. As for the classification, we use support vector machines (SVMs). In Section 2 we present a literature review on expressive interactions from computational aspects of speech and gesticulation. Then we describe our multimodal data collection process and extent of the JESTKOD database. In Section 3 we present the proposed statistical model for agreement/disagreement classification in dyadic interactions. Experimental evaluations are given in Section 4. Finally in Section 5, we provide conclusions.

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

2. MULTIMODAL DYADIC INTERACTION DATABASE

2.1. Literature Review

There are a variety of multimodal databases that contain continuous affect annotations and made publicly available for research purposes. The SEMAINE database consists of audio-visual data in the form of conversations between participants and a number of virtual characters with particular personalities [11]. The HUMAINE database includes a large collection of multimodal naturalistic and induced recordings [12]. We note that, although motion capture technologies are becoming widely available, there exist only a limited number of audio-visual databases which also include 3D motion data for modeling bodily gestures. In [13], Heloir et al. explore technical setups, scenarios and challenges in building a motion capture database for virtual human animation. Busso et al. present their interactive emotional dyadic motion capture (the USC IEMOCAP) database in [14], which is a multimodal and multi-speaker database of improvised dyadic interactions. The USC CreativeIT database contains full-body motion capture information in the context of expressive theatrical improvisations [15]. The database is annotated in the valence-activation-dominance space as well as the theater performance ratings such as interest, naturalness. Since interaction performances of the CreativeIT database are theatrical, speech and body gestures are rather amplified and pretentious in this database.

2.2. JESTKOD Database

Our main motivation to construct the JESTKOD database is to address more natural and affective dyadic interactions that can provide a valuable asset for investigating gesture and speech signals [16]. The JESTKOD database consists of dyadic interaction recordings of 10 participants (4 female/6 male, ages from 20 to 25) collected in 5 sessions, all in Turkish. In each session, 13-17 different topics exist in which both participants agree and disagree, and each conversation is around 5 minutes. Participants are recorded by a high-definition video recorder, full body motion capture system with 120 fps and individual audio recorders (44.1 kHz sampling rate, 16 bit, mono). A scene of a session which participants wear special suit with infrared markers for motion capture system is shown in Fig. 1. Full body motion capture is executed using the *Motive* software.

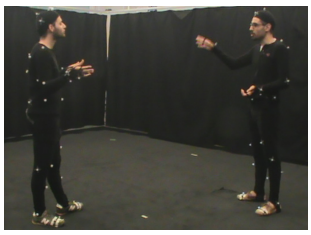

Fig. 1: A sample from video recordings of the JESTKOD.

Topics of the dyadic interactions are set by the moderator of the session using a preliminary information form. This form is given to the participants before the recordings to collect information on their likes and dislikes on common topics as sports, movies, music, etc. A summary of the topics is given in Table 1. In the JESTKOD database we have 66 dyadic interactions in agreement, and 79 dyadic interactions in disagreement.

Table 1: Summary of topics in the JESTKOD database

Pair #	Topics in the JESTKOD database			
	Agreement scenario	Num. clips	Disagreement scenario	Num. clips
1	Cinema, World cuisine, Holiday resorts, TV series	13	Football, Maths, Game consoles, PC Games	13
2	Football, World cuisine, Music, Cinema, Literature	13	Geography, Holiday resorts, PC Games, Theatre, Dance	16
3	Cinema, Sports, PC Games, Music, World cuisine	11	Cinema History, TV series, Animals, Education	17
4	World cuisine, Holiday resorts, Science-fiction, History, Theatre, Cities	16	Football, Cinema PC Games, TV series, Literature, Physics	17
5	Cinema, Languages, PC Games, Cities, Game consoles	13	Cinema, Sports, Holiday resorts, Nutrition, Musicals	16
Total		66		79

3. AGREEMENT/DISAGREEMENT CLASSIFICATION

We pose a two-class dyadic interaction type (DIT) estimation problem of agreement and disagreement classes from speech and motion modalities. A block diagram of the classification system is given in Fig. 2. Speech and motion streams of the two participants of the dyadic interaction are the inputs of the feature extraction block. Then frame level features of speech and motion goes into utterance extraction block to compose temporal collection of feature vectors. Joint and split speaker models perform a statistical or i-vector based 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.

3.1. Feature and Utterance Extraction

We utilize widely used feature representations for speech and arm motion modalities. Speech signal of the i-th participant, $S_i(t)$, 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_i}. On the other hand frame level motion feature vector f^{M_i} for the i-th 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 differ and they are respectively set as 100 and 120 fps.

In the utterance extraction we collect frame level feature vectors over the temporal duration of the utterance and construct matrices of features. We filter out the silence frames for the speech modal-

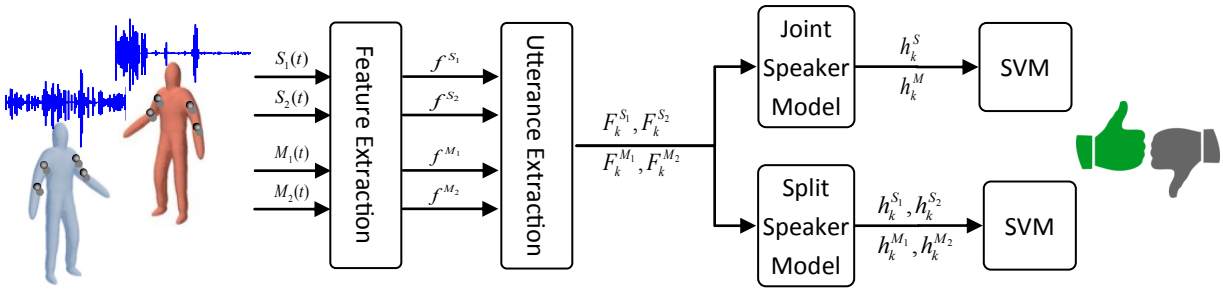

Fig. 2: Block diagram of the agreement/disagreement classification system.

ity and construct speech feature matrix as $F_k^{S_i} = [f_1^{S_i} \cdots f_{N_S}^{S_i}]$ for the k-th utterance with dimensions $39 \times N_S$. Similarly the motion feature matrix is constructed as $F_k^{M_i} = [f_1^{M_i} \cdots f_{N_M}^{M_i}]$ for the k-th utterance with dimensions $24 \times N_M$ without silence filtering.

3.2. Feature Summarization

We perform two summarization schemes, statistical functionals and i-vector representation to map the high dimensional utterance level feature matrices to low dimensional feature representations.

We use statistical functionals 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 as defined in [17].

The i-vector representation in total variability space (TVS) is an alternative to summarize the temporal features which is widely used in language and speaker recognition [18], as well as recently applied in age estimation [19] and discrete and continuous emotion classification [20, 21]. In the TVS, first a Gaussian mixture model is used to model the distribution of the data as:

$$P(F) = \sum_{i=1}^{L} \omega_i \mathcal{N}(F; \boldsymbol{\mu_i}, \sigma_i), \tag{1}$$

where F is the feature space, L is the total number of mixtures, ω_i, $\boldsymbol{\mu_i}$, and σ_i are the weight, mean vector, and covariance matrix of the i-th Gaussian mixture respectively. Then the super-vector, which is the concatenation of mean vectors $\boldsymbol{\mu_i}$, is mapped to a lower dimensional TVS space as, $\boldsymbol{\mu} = \boldsymbol{m} + T\boldsymbol{w}$, where $\boldsymbol{\mu}$ is the super-vector, $\boldsymbol{m}$ is a representative, usually the concatenated mean vectors of the universal background model (UBM), T matrix represents the TVS basis, and $\boldsymbol{w}$ is the extracted feature vector, which is known as i-vector in verification literature. The details to calculate T matrix are given in [18].

Using the statistical functionals and i-vector representations we construct two models, joint and split speaker models, for the two class DIT estimation. These two models are described in the following subsections.

3.2.1. Joint Speaker Model

In the joint speaker model (JSM) we collect features of participants together and then apply statistical or i-vector based summarization. Hence for the speech modality the combined features $[F_k^{S_1} F_k^{S_2}]$ are fed into statistical or i-vector based summarization to extract summarized feature representation h_k^S. Similarly, summarized feature representation h_k^M for the motion modality is extracted from $[F_k^{M_1} F_k^{M_2}]$ using statistical or i-vector based summarization.

3.2.2. Split Speaker Model

In the split speaker model (SSM) we apply statistical or i-vector based summarization for each participant and then combine the summarized features to represent speech and motion modality. Hence, based on statistical or i-vector summarization, summarized feature representations $h_k^{S_i}$ and $h_k^{M_i}$ are extracted from $F_k^{S_i}$ and $F_k^{M_i}$ for the speech and motion modalities, respectively. Then, the combination of $h_k^{S_1}$ and $h_k^{S_2}$ is used as the summarized feature representation of the speech modality. 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

We use support vector machine (SVM) as a binary classifier for the DIT estimation. Let us define a notation to describe an SVM classifier using feature vector h as SVM(h). Then we can define unimodal and multimodal classification tasks under JSM as SVM(h^S), SVM(h^M) and SVM(h^S, h^M) respectively for the speech, motion and multimodal classifiers. Similar unimodal and multimodal classification tasks under SSM can be defined as SVM(h^{S_1}, h^{S_2}), SVM(h^{M_1}, h^{M_2}) and SVM($h^{S_1}, h^{S_2}, h^{S_1}, h^{S_2}$).

4. EXPERIMENTAL EVALUATIONS

In the JESTKOD database, we have 66 positive and 79 negative dyadic interaction clips. We use leave-one-clip-out training, where we test one recording at a time and train models on the remaining recordings. We adjust the PCA output dimension to preserve 90% of the total variance for the output of statistical function. We use 128 Gaussian mixtures with diagonal covariance for TVS and 10 expectation-maximization iterations for the extraction of UBM and T matrix. We employ 30 dimensional i-vectors. The MSR Identity Toolbox [22] is used for UBM and TVS calculation. We use SVM with linear kernel from LibSVM package [23] and calculate the performance as the average of agreement and disagreement classification accuracy. The chance level recognition rate on the database is calculated as 49.99%. We performed classification evaluations at clip level and at utterance level. In the following two subsections we describe them separately.

Table 2: Unimodal and multimodal classification accuracy for clip level DIT estimation

method	Accuracy
JSM: i-vector(Motion)	55.74%
JSM: i-vector(Speech)	99.18%
JSM: i-vector(Speech+Motion)	98.36%
SSM: i-vector(Motion)	57.38%
SSM: i-vector(Speech)	85.25%
SSM: i-vector(Speech+Motion)	86.89%
JSM: statistics(Motion)	82.79%
JSM: statistics(Speech)	83.61%
JSM: statistics(Speech+Motion)	86.07%
SSM: statistics(Motion)	79.51%
SSM: statistics(Speech)	89.34%
SSM: statistics(Speech+Motion)	90.16%

4.1. Clip Level Classification

In clip level classification, we concatenate and summarize the features of a clip and estimate the DIT per clip. Classification accuracy results of four experiments, JSM with i-vector, SSM with i-vector, JSM with statistics, and SSM with statistics, with unimodal and multimodal classifier are presented in Table 2.

In all experiments, classification results of motion have the lowest accuracy. Moreover, i-vector works worse than statistical features for motion. Since the variation of motion is higher than MFCC and the data set is limited, the i-vector and GMM modeling suffer to create a better model. In all experiments except utilizing the JSM with i-vector, which has the lowest motion accuracy, the multimodal scenario has the highest accuracy as expected.

Comparing the results of SSM and JSM models, JSM does not work well with statistical functionals since JSM has a combinations of features from two participants. However, using GMM for summarization with i-vector representation in JSM delivers a better modeling. Moreover, SSM does not work well for i-vector since it doubles the feature size.

4.2. Utterance Level Classification

We also investigate effects of the utterance duration on the classification accuracy for DIT estimation. In the utterance level classification we run the DIT estimation over overlapping utterances. We use the JSM with i-vector and SSM with statistical functionals, which have the highest clip level results, over varying utterance durations. Here the duration is the total time of dyadic interaction, including silent and speech segments.

Classification accuracy of the two mentioned experiments as a function of utterance durations in the range [5,100] sec with step of 5 sec are depicted in Fig. 3 and Fig. 4. Note that when utterance duration is larger than 15 sec, the multimodal accuracy is higher than 80% for the binary agreement and disagreement classification task. For the SSM with statistical functionals when the duration is greater than 75 sec accuracy reaches to the clip level accuracy, which is around 90%. Furthermore multimodal performance curve always has the highest accuracy. However for the JSM with i-vector, motion modality performs poorly, and the speech and multimodal performance curves result similar to each other.

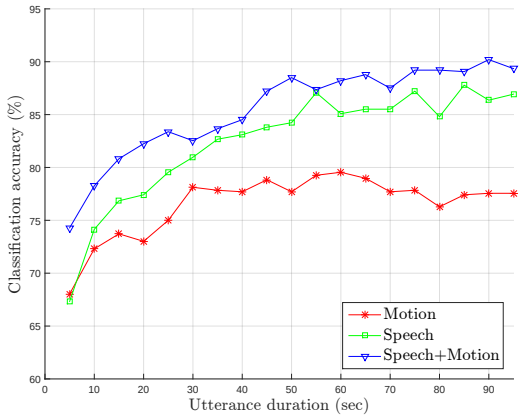

Fig. 3: Average agreement/disagreement classification accuracy as a function of utterance duration with SSM and statistical functionals.

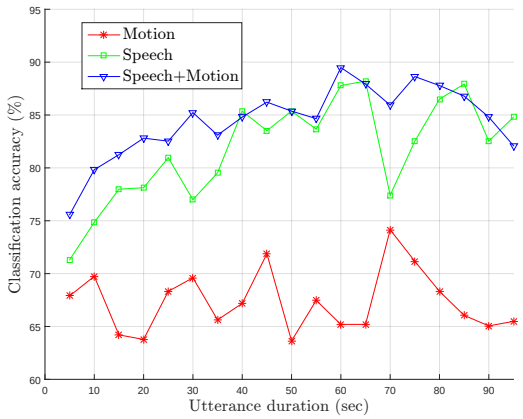

Fig. 4: Average agreement and disagreement classification accuracy as a function of utterance duration with JSM and i-vector.

5. CONCLUSIONS

In this paper we introduced the multimodal JESTKOD database and presented early results on agreement/disagreement classification of dyadic interactions over low level speech and motion capture representations. JESTKOD is a multimodal database of speech, motion capture and video recordings of affective dyadic interactions. We trained and tested SVM binary classifier using unimodal and multimodal features from speech and motion data to classify the interaction as agreement or disagreement. We provide the joint and split speaker model to estimate the dyadic interaction type by utilizing i-vector or statistical functionals to summarize the temporal features. Our findings suggest that the low level speech features carries more discriminative clues than the motion features. However, the multimodal features increases the accuracy of DIT classification with the statistical functions. Hence low level motion features carry additional information to discriminate agreement and disagreement given speech features.

6. 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 Tran. on*, vol. 3, no. 1, pp. 69–87, Jan 2012.

[2] 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 Tran. on*, vol. 27, no. 3, pp. 305–317, March 2005.

[3] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. of the 2003 Con. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proc. of HLT-NAACL 2003–short Papers - Volume 2*, Stroudsburg, PA, USA, 2003, NAACL-Short '03, pp. 34–36, Association for Computational Linguistics.

[4] 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 *Proc. of the 42Nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004, ACL '04, Association for Computational Linguistics.

[5] 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 Con. on*, March 2012, pp. 5089–5092.

[6] 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 Con. on*, March 2011, pp. 746–752.

[7] Wen 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 Con. on*, May 2011, pp. 5556–5559.

[8] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Con. on*, Sept 2009, pp. 1–9.

[9] J. B. Bavelas, N. Chovil, L. Coates, and L. Roe, "Gestures specialized for dialogue," *Personality and social psychology bulletin*, vol. 21, no. 4, pp. 394–405, 1995.

[10] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan, "Analysis of interaction attitudes using data-driven hand gesture phrases," in *Proc. of IEEE International Con. on Audio, Speech and Signal Processing (ICASSP)*, May 2014.

[11] 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 Tran. on*, vol. 3, no. 1, pp. 5–17, Jan 2012.

[12] 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*, A.R. Paiva, R. Prada, and R. Picard, Eds., vol. 4738 of *Lecture Notes in Computer Science*, pp. 488–500. Springer Berlin Heidelberg, 2007.

[13] Alexis Heloir, Michael Neff, and Michael 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.

[14] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008.

[15] 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.

[16] 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 Con. (SIU), 2015 23th*. IEEE, 2015, pp. 1374–1377.

[17] 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.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Tran. on*, vol. 19, no. 4, pp. 788–798, 2011.

[19] M. H. Bahari, M. McLaren, H. V hamme, and D. A. van Leeuwen, "Age estimation from telephone speech using i-vectors," in *INTERSPEECH 2012, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 506–509.

[20] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," in *Thirteenth Annual Con. of the International Speech Communication Association*, 2012.

[21] H. Khaki and E. Erzin, "Continuous emotion tracking using total variability space," in *Sixteenth Annual Con. of the International Speech Communication Association*, 2015.

[22] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker recognition research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[23] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Tran. on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.