On Head Motion for Recognizing Aggression and Negative Affect during Speaking and Listening

Fitrianie, Siska; Lefter, Iulia

DOI
10.1145/3577190.3614169

Publication date
2023

Document Version
Final published version

Published in
ICMI 2023 - Proceedings of the 25th International Conference on Multimodal Interaction

On Head Motion for Recognizing Aggression and Negative Affect during Speaking and Listening

Siska Fitrianie
Delft University of Technology
Delft, The Netherlands
s.fitrianie@tudelft.nl

Iulia Lefter
Delft University of Technology
Delft, The Netherlands
i.lefter@tudelft.nl

ABSTRACT

Affective aggression is a form of aggression characterized by impulsive reactions driven by strong negative emotions. Despite the extensive research in the area of automatic emotion recognition, affective aggression is a phenomenon that has received less attention. This study investigates the use of head motion as a potential indicator of affective aggression and negative affect. It provides an analysis of head movement patterns associated with various levels of aggression, valence, arousal and dominance, and compares behaviors and recognition performance under speaking and listening conditions. The study was conducted on the Negative Affect and Aggression database - a multimodal corpus of dyadic interactions between aggression regulation training actors and non-actors, annotated for levels of aggression, valence, arousal, and dominance. Results demonstrate that head motion features can serve as promising indicators of affect during both speaking and listening. Valence and arousal prediction achieved better performance during speaking, while aggression and dominance were better predicted during listening. Significant increases in the magnitude of pitch angular acceleration were associated with escalation along all four annotated dimensions. Interestingly, higher escalation was accompanied by a significant increase in the total number of movements during speaking, but a significant decrease of the number of movements was observed as escalation increased along listening intervals. These findings are particularly relevant as head motion can be used solely or potentially as a supplementary modality when other modalities such as speech or facial expressions are unavailable or altered.

KEYWORDS

Affective Computing, Head Motion, Affective Aggression, Emotion, NAA database, Speaking and Listening Behavior.

ACM Reference Format:
Siska Fitrianie and Iulia Lefter. 2023. On Head Motion for Recognizing Aggression and Negative Affect during Speaking and Listening. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614169

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614169

1 INTRODUCTION

The past decades have witnessed significant progress in the area of automatic affect recognition, with various modalities such as facial expressions, speech, and bodily gestures being used as expressive signals [51]. While studies targeting recognizing emotions modelled as either discrete or continuous were initially most prominent [69], interest has grown in identifying behaviors and states related to medical conditions, such as depression [18], anxiety [56], and pain [20]. However, one category of behaviors that has received relatively little attention thus far, and which is related to affective states and social interaction, is those associated with affective aggression.

Human aggression is any behavior directed toward another individual that is carried out with the immediate intent to cause harm [6]. Aggression has numerous negative consequences to victims, witnesses, as well as to perpetrators, including the risk of emotional, behavioral and mental-health problems. In terms of form and function, two types of aggression are being distinguished in literature, namely affective (reactive) and instrumental (proactive) aggression. While instrumental aggression is represented by purposeful and planned actions to achieve a goal and is associated with covert behaviors, affective aggression results from impulsivity and negative emotions in reaction to a threat or provocation, as in the case of a heated argument, and is associated with high sympathetic activation and overt behavior [16][11]. Recognizing levels of (imminent) aggression has important applications, including the development of mental health therapies for aggression-related disorders, self-management, surveillance, and improving human-agent interaction.

While head motion has traditionally been considered a nuisance and filtered out [50] in affect recognition, there is growing evidence that it can provide valuable information for recognizing a persons's affective states [1][2][24][57][65][66] and personality traits [46]. Head motion analysis brings several advantages, being more robust than other signals. As opposed to facial expressions, which can be altered by speech or viewing angle, and speech, which is only available when people are speaking, head motion is an available signal both during speaking and listening. The analysis of head motion is furthermore appealing in the case of Virtual Reality (VR)-related applications [66] since head-mounted displays contain sensors that track head movement (gyroscopes) and can seamlessly be incorporated in analysis. In particular, in the context of aggression, VR-enabled aggression regulation training solutions are emerging with the promise of allowing subjects to develop skills in a safe and controlled environment, without exposure to real threats, as for example treating forensic psychiatric inpatients [22] and training clinicians exposed to aggression to de-escalate [48].

This paper investigates whether affective aggression, hereinafter referred to as aggression, as well as negative affective states, represented as levels of valence, arousal, and dominance, can be predicted based on head motion. Furthermore, the paper provides an analysis of head movement patterns associated with levels of aggression and negative affect while a person is either speaking or listening, as well as prediction performance under these two conditions. The latter is motivated by the fact that movements of the head encompass both affect [30] and discourse [47] related information, which can lead to differences in head motion patterns and recognition performance during speaking and listening behaviors. The experiments are performed on the NAA dataset of negative affect and aggression [45], a corpus of improvised face-to-face interactions between actors specialized in aggression regulation training and non-actor participants.

The next section provides an overview of related work on aggression recognition and head-motion-based affective states prediction. Then, the NAA dataset is described in section 3. The paper continues with the feature extraction procedure in section 4 and methodology in section 5, followed by results and discussion in section 6 and finally conclusions.

2 RELATED WORK

This section begins with an overview of studies related to the automatic recognition of aggression and negative affective states based on multiple modalities and then zooms in on studies on affect recognition using head motion.

2.1 Automatic Recognition of Aggression and Negative Affect

Aggressive behaviors are categorized into two sub-types: instrumental aggression (proactive) and affective aggression (reactive). Instrumental aggression is considered "cold-blooded" as it is motivated by well-calculated intentions of harm and associated with low emotional arousal. Conversely, affective aggression is linked to strong negative emotions and a lack of impulse control and is therefore considered "hot-blooded" [16][5]. Its manifestations include various verbal and non-verbal behaviors and interactions, which can be picked up by behavior recognition systems.

Numerous studies investigated automatic violence detection in the field of computer vision [54], focused on recognizing actions depicting violence such as throwing and kicking based on video or kinematic (3-D) and electromyographic performance data [64], or extracted aggression from audio-visual inputs [68]. It is important to note that while focusing on aggression, the interaction in this study are more subtle and do not include violence, which is generally regarded as an extreme form of physical aggression [3].

While the distinct topic of (affective) aggression was less explored by the research community compared to emotion recognition in general, studies on negative affective states and problematic interactions are related to our work. The speech modality was leveraged to automatically recognize escalating negative interactions [40], to recognize manifestations of anger in call centers [55], and to detect frustration elicited by playing a game [61]. Several characteristics of problematic interactions were identified, such as decreasing inter-personal space [17], and the use of gestures with specific meaning and movement modulation [41]. An interesting finding related to speech was that aggression [44] and conflicts in political debates [35] were associated with a high degree of overlapping speech. Furthermore, including speech overlap as a feature proved to be one of the most discriminative features in these cases.

2.2 The Role of Head Motion in Affect Recognition

While it may not be the most prominent signal one would perceive in non-verbal communication, head motion is a very rich signal. On the one hand, it conveys discourse-related information [47]. While stillness tends to occur during pauses and while listening, it marks the structure of the ongoing discourse and is used to regulate interaction [26]. On the other hand, head motion communicates a variety of feelings, thoughts, and emotions. For instance, a bowed head connotes submission and inferiority emotions, such as shame, shyness, regret, guilt, and embarrassment [4]. If the chin is lifted it connotes dominance, superiority emotions (even arrogance), joy, and contentment, while a lowered chin indicates a negative or aggressive attitude [53]. Clinical patients with depression were found to move their head less and slower and postured it more downwards compared to healthy controls or the same subjects after successful treatment [2]. People in pain were associated with head movements and postures that tend to be oriented downwards or towards the pain site and involve a high movement range and faster movement of the head in painful situations [65].

Concerning automatic analysis and recognition of human affective states, research showed the significance of head motion alone in conveying affect [1] [2][24][57][65]. In particular, Adams et al. [1] and Samanta & Guha [57] found that head motion carries complementary information to conversational facial expressions. The combination of features was also used to categorize mental states or a subset of communicative labels [21][63][25], whereas Silva & Bianchi-Berthouze [59] combined them with body posture features to classify emotional expressions. Other research used the results of automatic head-tracking and analysis in face-to-face interaction, i.e., between an avatar and an adult human [13], and between distressed couples [28]. Hammal et al. [29][30] analyzed head movements during positive and negative affective states of infants interacting with their mothers, while Madan et al. [46] used head features for personality trait recognition.

Different techniques were used for data acquisition in automatic emotion recognition based on head pose and motion. For example, Gunes & Pantic [24] used the magnitude and direction of 2D head motion and head gestures, such as nods and shakes, to predict emotions in a continuous dimensional space. Other studies used video data and computer vision techniques to estimate 3D directions of head poses automatically [49]. While other studies approach head motion of individual persons, Tan et al. focused on estimating joint head orientation of interacting group members as a cue to social attention [62].

In terms of differences between negative and positive affect, Hammal et al. [29][30] showed that angular velocity and acceleration of pitch, yaw, and roll were higher for negative than for positive affect in infants; whereas in [28], they showed that pitch angular displacement was higher during conflict than during neutral states.

Sitting individual exposed to visual emotional stimuli were found to lean more forward and moved more when watching positive stimuli than when watching negative stimuli [9].

While several efforts focused on using head motion for recognizing negative affective states, e.g., anger, fear, pain, stress, and depression, [1][2][23][59][65], to the best of our knowledge this is the first attempt to use head motion to predict aggression. Another distinct feature of this work is that we distinguish between speaking and listening conditions.

3 THE DATASET OF NEGATIVE AFFECT AND AGGRESSION

The experiments were performed on the NAA Dataset of Negative Affect and Aggression [45]. The dataset was collected in the context of a project investigating the effectiveness of Virtual Reality (VR) aggression prevention therapy in the case of inpatients in psychiatric forensic clinics [22]. The dataset consists of dyadic face-to-face interactions (role-plays) between professional aggression regulation training actors and non-actors. These role-plays were designed by psychologists and psychiatrists and are similar to the ones practiced in therapy. To ensure realism, the only instruction the participants received were short role descriptions, including the context (involving a degree of urgency), their role (e.g., bus driver), and their goal (e.g., do not let the passenger travel without paying). Since no scripts were used, the interactions emerged as the participants reacted to one another, and could, therefore best be described as improvisations. Besides the degree of urgency in the specified scenarios, the actors played an important role in manipulating the emotional content of the database. Being proficient in giving aggression management training in forensic clinics and in showing and eliciting emotions, they were instructed to vary the degree of aggression for different scenarios.

As it can be considered that the actors had the role of an emotional stimulus (they were displaying various levels of aggression towards the non-actors), previous work performed a validation study to check whether the interactions resulted in the emotional arousal of non-actors [42]. The analysis found the relationship between how the heart rate variability (as an indication of sympathetic activation) of non-actors varied in response to the varying aggression levels portrayed by the actors. While the study makes no claims about specific experienced emotions, the results indicate that the behavior of the actors was indeed able to stimulate the non-actors.

In total, 16 subjects participated in the recordings: 4 actors (3 male, 1 female) and 12 non-actors (3 male, 9 female). Three different scenarios were played multiple times according to the following protocol. Each actor played every scenario thrice, each with a different aggression level evolution (escalate, de-escalate, keep aggression high) and with a different non-actor. Each non-actor participated in each of the three scenarios once, every time with a different actor, and every time with a different aggression level variation. The scheme led to 36 recorded interactions of approximately 3 minutes each. Multiple sensors were used for recordings, including microphones, cameras, MS Kinect, physiological sensors, as well as gyroscopes for tracking head motion, the latter being used in this study.

Table 1: Inter-rater agreement (Weighted Krippendorff's alpha with quadratic weights [37]) of label annotation for different conditions as reported in [45].

	A	NA	A-S	NA-S	A-L	NA-L	S	L
Agg	.79	.45	.81	.48	.74	.30	.79	.71
V	.77	.62	.77	.62	.75	.62	.72	.67
A	.71	.38	.72	.40	.67	.30	.68	.60
D	.62	.50	.65	.51	.52	.42	.58	.46

Note: Agg = aggression; V = valence; A = arousal; D = dominance; A = actor; NA = non-actor; A-S = actors, speaking condition; NA-S = non-actors, speaking non-actors; A-L = actors, listening condition; NA-L = non-actors, listening condition; S = all speech segments; and L = all listening segments.

The recordings were manually segmented into utterances based on turn-taking and by splitting longer utterances at pauses, which resulted in 2420 utterances (duration $M = 2.56s$, $SD = 1.7$). Each dimension was annotated by 3 raters. Degrees of Aggression, Arousal, and Dominance were annotated on a 5-point scale. As an exception, Valence was annotated on a 9-point scale to ensure a 5-point granularity for negative Valence. The annotation was performed in turn per dimension and per participant in interaction, meaning that for every utterance there is a label for the actor and a label for the non-actor.

For this study, the annotation was reduced to a 3-point scale, corresponding to low, medium, and high levels of each annotated dimension, as some of the classes were very little represented. The label mapping procedure has a high impact on the experiments, and the procedure presented has been empirically chosen based on initial experimentation and class balance considerations. Some classes (and these differ per annotated dimension) are little represented. Examining the confusion matrices between annotators we observed that most confusions were between neighbouring classes. As part of the re-labelling procedure we experimented with multiple options for combining neighbouring classes together to obtain a better representation of each class. These options maintain the categories of low, medium and high, even though the cutting points might vary. Full experiments (just as the ones described in the paper) were performed with the other setups as well, and hereby we present results of the best performing relabeling procedure.

For Aggression, the two lowest levels were merged into a new class 1 (low aggression), the middle class became the new class 2 (medium aggression), and the two highest levels were merged into a new class 3 (high aggression). In the case of Valence, the two most negative levels were merged into a new class 3 (very negative), the next two levels were merged into a new class 2 (moderate negative), and the remaining levels were merged into a new class 1 (non-negative valence). For Arousal, the lowest level was considered as a new class 1 (low arousal), the next two increasing levels were merged into a new class 2 (medium arousal) and the two highest levels were merged into a new class 3 (high arousal). For Dominance, the lowest two levels were clustered into a new class 1 (submissive), the middle level formed a new class 2 (moderate dominance), and the two highest levels were merged into a new class 3 (very dominant).

The inter-rater agreement was computed for the whole dataset, but also separately for actors, non-actors, and segments when the participants were speaking and listening, as shown in Table 1. The values presented in the table are different from the ones reported in [45], where Krippendorff's alpha with linear weights was utilized as measurement. Given that by observing confusion matrices between raters, confusion between the neighboring classes was noticed, herein we report agreement using weighted Krippendorff's alpha with quadratic weights. This method is suitable for ordinal data as it penalizes confusions between distant levels more and between neighboring levels less [7]. The agreement for rating the actors is consistently higher than for the non-actors for each annotated dimension. The main reason for this discrepancy might be that the behavior of the actors was much more overt (e.g., ample gesturing, loud voice), while most non-actors had a much subtler behavior. In addition, the results show that the inter-rater agreement was higher for segments that include speech and lower for segments that include silence (listening) for both actors and non-actors. This finding indicates that non-verbal behavior is harder to interpret when observing participants who are listening. The distinctions between speaking and listening behavior are explored in this paper in terms of its influence on aggression/affect recognition from head motion and speech.

4 FEATURE EXTRACTION

Head motion data was recorded using gyroscopes (Chrum, based on CH-Robotics UM-6 IMU) attached to the head of each participant using a head band. The raw data were segmented following the utterance-based segmentation. Therefore, each segment corresponds to an aggression, valence, arousal, and dominance level.

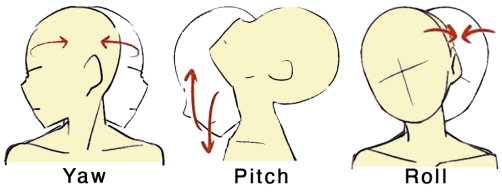

Figure 1: Head orientation angles: pitch, yaw, and roll

At the beginning of each scenario, the participants were facing each other for calibration. Head rotations around the three axes known as pitch (i.e., looking up- or downward), yaw (i.e., looking to the left- or right), and roll (i.e., tilting clockwise or anticlockwise) (see Figure 1) were used as a basis for feature extraction. This results in features being extracted over 6 directions of movement.

Inspired by previous studies [2][57][28][29][65], we used the following procedure for feature extraction from the raw head tracker data. First, we considered an interval of -5 to 5 degrees of pitch, yaw, and roll as looking forward (keeping into account the initial calibration). For measuring changes in head movement within a segment, we converted the head angles to angular displacement, angular velocity, and angular acceleration. The angular displacements of pitch, yaw, and roll were computed by subtracting the overall mean head angle from each observed head angle, whereas angular velocity and angular acceleration were computed as the derivative of angular displacement and angular velocity, respectively. Further, to measure the magnitude of variation of angular displacement, angular velocity and angular acceleration, the root mean square (RMS) of these variables was computed.

For each consecutive segment and each actor and non-actor data, 117 features were computed separately (corresponding codes are available online [43]). These are:

(a) The maximum, minimum, mean, variance, standard deviation, range, and RMS values of angular displacement, angular velocity, and angular acceleration quantities (7 x 9 features);
(b) The maximum, minimum, range, mean, and average duration values of: looking left, right, up, and down, and tilting clockwise and anticlockwise (4 x 6 features); Besides looking forward, the head can move to these six directions.
(c) The average duration of the time intervals when the head moved from looking forward to the other directions (6 features) and, conversely, from the other directions to looking forward (6 features);
(d) The average duration of the time intervals that the subject was looking forward, within a segment (1 feature);
(e) The ratio of movements in each of the six directions (the number of movements in each direction divided by the total number of movements) within a segment (6 features);
(f) The number of the head movements from looking forward to the six directions (6 features) and the sum of the number of these movements (later referred to as total movements) (1 feature);
(g) The number of looking direction changes to the opposite side (i.e., zero-crossing rates or Z-cross e.g., left to right) and the total number of these changes (4 features) .

5 METHODOLOGY

The experiment explores the recognition performance of Aggression, Valence, Arousal, and Dominance while the participants were either only speaking or only listening. Furthermore, trends of the most important features in the recognition process were analyzed. The corresponding codes and analysis results are available as supplementary material [43].

5.1 Data Partitioning

For this experiment, samples from all participants (actors and non-actors) were separated into two categories: speaking and listening. They were used to develop two models: (a) a Speaking model, trained on data where the participants were speaking (3146 samples), and (b) a Listening model, trained on data where the participants were listening (1688 samples). It is interesting to note that the data contains a high amount of overlapping speech (1458 samples) [44], which means that both the actor and the non-actor were speaking simultaneously during those segments. The overlapping speech segments were included in the Speaking model: for each overlapping speech segment, the head features of the actor were mapped on the label of the actor for that segment, and similarly for the non-actor. These segments were excluded from the Listening model.

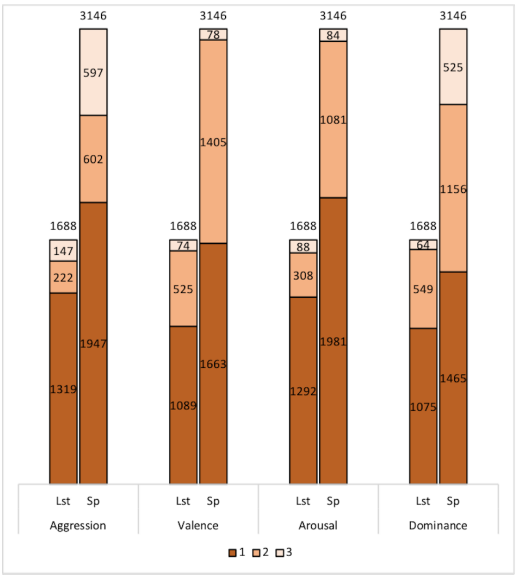

Figure 2: The number of samples per class for the four annotated dimensions used for the Listening (Lst) and Speaking (Sp) models

5.2 Classification

To ensure person independence, the experiments were performed using a leave-one-subject-out (LOSO) cross-validation. The classification was performed using five different classifiers: Logistic Regression (LR) [12], Support Vector Machine [10], Random Forest (RF) [14], Balanced Random Forest (BRF) [14], and Conditional Random Fields (CRF) [39]. For the last classifier, the classification experiments were performed using the standard implementations from the `Sklearn-crfsuite` package, while the others were from the `Scikit-learn` package.

Before classification using LR, SVM, RF, and BRF, to mitigate the effects caused by data being unevenly distributed over the three degrees of aggression, valence, arousal, and dominance, as shown in Figure 2, statistical minority oversampling (SMOTE) with Tomek Links [8] was applied. This method generates new artificial samples of the minority class by adding noise to the data. Based on the initial label distribution of the training set per fold, statistical oversampling was applied for the two least represented classes, with a pre-computed parameter to even out the distributions. SMOTE-Tomek Links was applied only on the training set for each fold. Before classification, the features were normalized by scaling them into the range of 0 to 1.

In contrast to the other classifiers explored, CRF is the only approach that takes into account the temporal dimension: past states will influence the current state. Hence, segments were considered in the order in which they appeared in the recordings, and the temporal sequences were considered per participant in the interaction. However, it was not assumed that the observations (i.e., segments) were conditionally independent given the hidden states. Therefore, given a sequence of segments, $X = \{x_1, x_2, ..., x_t\}$ where each x_i is a feature vector, CRF was applied to specify the conditional probability $P(Y|X)$ of a label sequence, $Y = \{y_1, y_2, ..., y_t\}$ and by using contextual information from previous labels, to predict the sequence. As a note: since CRF does not suffer from label bias problems and does global probability normalization, no statistical oversampling and normalization before classification using this method was applied.

5.3 Data Analysis

Recognition results are reported as unweighted average accuracy (UA), given the data imbalance. To check whether the differences between different conditions are significant, the pairwise comparison of recognition rates between the two models was calculated using a paired-samples t-test.

Besides comparing the overall results of different methods and conditions, we were interested in the discriminative value of individual features. The impact of features on prediction was assessed using permutation importance from the `rfpimp v.1.3.7` package. This method measures feature importance by observing the effect of randomly shuffling each feature on model accuracy [52]. The importance of a feature is measured by calculating the increase in the model's prediction error after permuting the feature. A feature is 'important' if shuffling its values increases the model error because the model relied on the feature for the prediction. A feature is 'unimportant' if shuffling its values leaves the model error unchanged because, in this case, the model ignored the feature for the prediction.

Furthermore, trends of specific features given changes in levels of Aggression, Valence, Arousal, and Dominance are explored. The tests were performed using the Kruskal-Wallis method since some continuous feature data were not normally distributed (results the Shapiro-Wilk tests can be found in the supplementary materials [43]). Kruskal-Wallis is a non-parametric method for comparing two or more independent samples of equal or different sample sizes [38]. This analysis provided reference data for further analysis between two levels of annotated dimension data. If the variance test statistic was significant ($p < .05$), then Dunn's non-parametric post-hoc tests were conducted to compare all of the combinations [19]. These tests were conducted using R-package `FSA v. 0.8.25`. Given space limitation, the feature importance and trends are only specified for the ten most important features.

6 RESULTS AND DISCUSSION

Table 2 presents the recognition rates for Listening (left) and Speaking (right) in terms of unweighted average accuracies. For the given three class problems, a Dummy Classifier that generates predictions uniformly at random for the current 3 class problem with unbalanced data achieved unweighted accuracies of 30-33% on the tasks. In terms of performance per task, Aggression and Arousal were recognized with higher accuracies (>80%) than Valence and Dominance. While all five classifiers performed higher than chance for all four task, the CRF classifier performed best for all annotated dimensions (except for Dominance during Speaking), followed by the tree-based classifiers. This may indicate the importance of taking the temporal aspect of the interactions into account.

Table 2: Average recognition rates (Unweighted Accuracy %) using Head features while Listening and Speaking.

	Listening				Speaking			
Classifier	**Agg**	**Val**	**Ar**	**Dom**	**Agg**	**Val**	**Ar**	**Dom**
Logistic Regression	58.51	42.12	46.82	49.91	59.91	43.91	50.27	47.77
Support Vector Machine	66.05	48.06	54.90	51.00	68.98	53.27*	66.59*	45.79
Random Forest	77.42	59.58	76.26	61.51*	73.11	63.56	76.53	48.28
Balanced Random Forest	78.30	60.43	75.88	64.30*	74.77	63.96	76.74	49.05
Conditional Random Fields	85.54*	64.11	81.73	69.90*	79.02	67.42	77.23	48.48

Note: $*$ significantly higher accuracy among the Listening and Speaking conditions, $p < .05$.

Analyzing the differences in performance between the Listening and Speaking conditions, it can be noticed (Table 2), with little exception, a consistent trend of higher recognition rates for Speaking in the case of Valence and Arousal (significant for SVM). In contrast, Aggression and Dominance were better predicted during Listening, with significant differences for CRF in the case of Aggression and RF, BRF, and CRF for Dominance.

Tables 3, 4, 5, and 6 present the ten most important features for Aggression, Valence, Arousal, and Dominance, respectively, ranked according to their permutation importance, together with the results of the one-way analysis of variance performed per dimension and speaking / listening condition and post-hoc analysis. Features derived from pitch (moving head up and down) are the most frequent in the feature importance table, accounting for 45% of the occurrences. Head yaw (looking to the left and right) related features account for 36% of occurrences, while roll (tilting head to the sides) is the least frequently represented, accounting for 9% of occurrences.

Several interesting trends are observed when analyzing the top ten most important head features and how they vary with increasing levels of the considered affective dimensions. The number of total movements and RMS of pitch angular acceleration were among the top predictors for all four dimensions and both for the Speaking and Listening conditions. While RMS acceleration of pitch consistently increased for all 4 dimensions accompanying escalation, both during speaking and listening, a different trend is observed for the number of total movements. Interestingly, increases in Aggression, negative Valence, Arousal, and Dominance were associated with a significant increase in total movements during Speaking, while a significant decrease in total movements was observed during Listening while all four dimensions were escalated.

For the Speaking condition, pitch angular acceleration significantly increased with the escalation levels of the four dimensions. A similar increasing trend was observed for the pitch angular velocity of the four dimensions, except for the lowest levels of Arousal. Further, significant increases were noted for several yaw-related features in Aggression, Valence, and Arousal and for roll in Aggression and Dominance.

For the Listening condition, although the variance of measurements based on pitch was lower than for Speaking, significant increases in angular acceleration and angular velocity were still reported in all dimensions as the levels escalated. This occurred from medium to high Arousal, between low and medium Aggression and Dominance, and from medium to high negative Valence. A similar trend was depicted by angular velocity and angular acceleration based on yaw and roll for Valence, Arousal, and Dominance. For Aggression, the angular displacement and velocity of yaw significantly decreased from medium to high Aggression, while roll angular acceleration significantly increased in lower Aggression. Measurements of several individual movements revealed significant decreases. For example, the average duration of looking forward to looking down between low and medium Dominance, the ratio of looking to the right for high Aggression, the average duration from looking forward to looking left for high negative Valence, and the average duration of changing from looking forward to tilting-clockwise between medium and high Arousal.

The analysis suggests that during speaking humans increase their head movements as aggression and negative affect are escalating, especially in the case of movements related to pitch and yaw. It seems that for these higher levels of escalation, the speakers move their heads, i.e., nod, bow, and shake, more often and more rapidly to reinforce their verbal messages, which is at the same time a display of their emotional states. This result supports a study investigating emotionally contrastive speech tasks that found increased head velocity under stress conditions [23]. Correspondingly, Hadar et al. [26] found a significant positive correlation between head movement amplitude and peak speech loudness, which was driven mostly by fast, high-intensity movements and loud sounds.

In the case of listening, the analysis of variance indicated that the higher the levels of escalation, the less movement there is, although the angular velocity and acceleration based on pitch and yaw are still increasing side-by-side with the levels of the annotated dimensions. This finding is in accordance with Hadar et al. [27], who found relatively little movement during listening turns. On the other hand, Yngve [67] and Duncan [32] regarded head motion while listening as backchannels. This may explain the increase of the angular velocity and acceleration on different levels of the four annotated dimensions (albeit having lower levels than during Speaking). Backchannels are considered the listeners' spontaneous expression when they desire to interject. They include those times when the listener is turning the head from side to side indicating disagreement [36], lifting the chin up to show dominance-superiority emotions [53], and increasing the pace of the movement to signal the lack of patience [53]. Moving one's head during listening may also be used to signal the desire to assume the role of speaker. According to Harrigan [31], listeners provide cues by increasing the head and gaze shifting as a nonverbal means of requesting a turn. Then, they will begin their turn by briefly looking away or averting

Table 3: One-way analysis of variance for Aggression levels during Listening and Speaking

	Listening					Speaking				
		Var	Post-hoc Analysis				Var	Post-hoc Analysis		
No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)	No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)
1	Total movements	132.46*	5.65*	10.64*	4.85*	1	RMS acc. pitch	493.33*	-15.16*	-19.35*	-3.43*
2	Ratio total look-up	8.18*	-2.84*	-0.02	1.92	2	Total movements	79.47*	-7.76*	-6.09*	1.33
3	Var velocity yaw	6.66*	-0.34	2.50*	2.27*	3	Var acc. pitch	499.99*	-15.18*	-19.53*	-3.56*
4	Range displacem. yaw	46.06*	1.47	6.75*	4.52*	4	Range velocity pitch	452.38*	-15.26*	-17.99*	-2.25*
5	Mean acc. roll	12.25*	-3.26*	-1.65	0.88	5	Ratio total look-right	2.94	-0.67	-1.69	-0.83
6	Var acc. pitch	16.78*	-3.71*	-2.17*	0.76	6	Total Z-Cross yaw	85.96*	-7.05*	-7.51*	-0.39
7	Range velocity pitch	10.37*	-3.18*	0.09	2.24*	7	Ratio total look-down	140.04*	-8.76*	-9.79*	-0.86
8	RMS acc. pitch	19.25*	-3.79*	-2.66*	0.41	8	Var velocity yaw	38.48*	-5.04*	-4.70*	0.26
9	Ratio total look-right	8.04*	-0.32	2.76*	2.47*	9	Max acc. pitch	464.96*	-14.58*	-18.88*	-3.52*
10	Max acc. pitch	7.12*	-2.66*	-0.55	1.36	10	Mean acc. roll	144.25*	-8.83*	-9.99*	-0.96

Note: $*$ significant $p < .05$, $df = 2$; Var = Analysis of Variance; RMS = root mean square

Table 4: One-way analysis of variance for Valence levels during Listening and Speaking

	Listening					Speaking				
		Var	Post-hoc Analysis				Var	Post-hoc Analysis		
No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)	No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)
1	Total movements	42.06*	1.73	4.22*	5.93*	1	RMS acc. pitch	315.98*	-2.60*	-7.90*	-17.03*
2	Mean dur. fwd. to left	19.21*	-1.15	0.75	4.35*	2	Mean acc. pitch	286.57*	-2.47*	-7.51*	-16.22*
3	Mean acc. pitch	23.24*	0.32	-1.75	-4.81*	3	Total movements	29.54*	-1.50	-3.04*	-4.95*
4	RMS acc. pitch	26.06*	0.16	-2.02	-5.07*	4	Mean dur. fwd.-right.	22.92*	0.86	-0.63	-4.78*
5	Mean velocity pitch	11.55*	0.26	-1.20	-3.39*	5	Var acc. pitch	315.71*	-2.57*	-7.87*	-17.03*
6	Var acc. pitch	23.59*	0.02	-2.04	-4.81*	6	Var acc. yaw	39.41*	-3.00*	-4.55*	-5.03*
7	Min displ. yaw	22.36*	-0.99	-2.85*	-4.43*	7	Mean dur. look-fwd	13.35*	1.64	2.57*	3.00*
8	RMS velocity pitch	13.48*	0.08	-1.48	-3.64*	8	Mean acc. yaw	45.88*	-3.06*	-4.78*	-5.54*
9	Var velocity pitch	12.97*	-0.01	-1.53	-3.56*	9	Min acc. yaw	0.07	-0.23	-0.26	-0.10
10	Min acc. yaw	17.15*	-0.04	-1.79	-4.09*	10	Min durat. look-down	36.18*	0.25	2.09	5.91*

Note: $*$ significant $p < .05$, $df = 2$; Var = Analysis of Variance; RMS = root mean square

Table 5: One-way analysis of variance for Arousal levels during Listening and Speaking

	Listening					Speaking				
		Var	Post-hoc Analysis				Var	Post-hoc Analysis		
No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)	No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)
1	Total movements	89.36*	2.43*	6.84*	8.81*	1	Total movements	108.75*	-0.42	-3.85*	-10.29*
2	Mean dur. fwd-tilt-CW	74.33*	1.77	5.91*	8.19*	2	RMS acc. pitch	467.00*	-2.96*	-9.89*	-20.90*
3	Min velocity roll	78.72*	-0.08	-4.69*	-8.81*	3	Var acc. pitch	469.38*	-2.90*	-9.85*	-20.97*
4	Mean velocity pitch	37.71*	-2.74*	-5.21*	-5.16*	4	SD acc. pitch	469.38*	-2.90*	-9.85*	-20.97*
5	Var velocity pitch	28.52*	-1.85	-4.19*	-4.77*	5	Mean acc. yaw	122.47*	-0.86	-4.47*	-10.85*
6	Mean acc. pitch	48.07*	-2.66*	-5.61*	-6.07*	6	Ratio total look-right	13.46*	1.64	0.46	-3.45*
7	RMS acc. pitch	45.81*	-2.23*	-5.24*	-6.10*	7	Mean acc. pitch	425.27*	-2.87*	-9.48*	-19.93*
8	Mean acc. yaw	21.73*	-0.18	-2.58*	-4.61*	8	Min velocity pitch	72.45*	0.08	-2.75*	-8.46*
9	RMS velocity pitch	34.00*	-2.21*	-4.70*	-5.12*	9	Ratio total look-down	147.91*	-1.25	-5.19*	-11.86*
10	Min displ. yaw	19.87*	-2.14*	-3.86*	-3.65*	10	Mean velocity pitch	325.41*	-2.67*	-8.43*	-17.39*

Note: $*$ significant $p < .05$, $df = 2$; Var = Analysis of Variance; CW = clockwise; RMS = root mean square

their gaze, which may include head shifting [33][34]. Nevertheless, the communicative cues during listening could also be recognized

Table 6: One-way analysis of variance for Dominance levels during Listening and Speaking

		Listening					Speaking				
		Var	Post-hoc Analysis				Var	Post-hoc Analysis		
No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)	No	Feature	χ^2	Z(1-2)	Z (2-3)	Z (1-3)
1	Total movements	97.08*	8.48*	6.13*	2.60*	1	RMS acc. pitch	446.71*	-13.35*	-20.47*	-10.85*
2	Mean dur. fwd.-down	119.41*	9.93*	5.88*	1.78	2	Mean acc. pitch	416.13*	-13.22*	-19.64*	-10.10*
3	Min displ. pitch	6.38*	-1.84	-1.97	-1.18	3	Var acc. pitch	441.58*	-13.04*	-20.42*	-11.04*
4	Mean acc. pitch	48.68*	-5.97*	-4.40*	-1.92	4	SD acc. pitch	441.58*	-13.04*	-20.42*	-11.04*
5	Mean acc. roll	32.22*	-5.43*	-2.38*	-0.16	5	SD velocity pitch	355.24*	-12.51*	-18.04*	-8.99*
6	Mean acc. yaw	27.11*	-5.20*	-0.98	1.11	6	Min velocity pitch	59.02*	-5.37*	-7.23*	-3.33*
7	Max acc. pitch	14.05*	-3.29*	-2.23*	-0.86	7	Ratio total look-down	107.82*	-6.12*	-10.18*	-5.80*
8	RMS acc. pitch	44.08*	-5.74*	-4.09*	-1.70	8	Total movements	22.66*	-2.03*	-4.76*	-3.36*
9	Var acc. pitch	36.48*	-5.26*	-3.66*	-1.48	9	Mean acc. roll	127.40*	-8.30*	-10.42*	-4.36*
10	RMS velocity pitch	26.93*	-4.77*	-2.69*	-0.73	10	RMS velocity pitch	339.16*	-12.50*	-17.51*	-8.44*

Note: $*$ significant $p < .05$, $df = 2$; Var = Analysis of Variance; RMS=root mean square

as signals for aggression and negative affective states. In the case of Dominance, the changes in angular velocity and acceleration occur significantly from the low to medium level.

6.1 Limitations

This study was performed on improvised interactions between professional aggression regulation training actors and non-actors and may suffer from well known artefacts resulting from the use of actors, see for example [15], [58]. Data streamed from head-attached sensors was analyzed and interpreted as head motion, without taking into account influences from moving one's whole body. This can be seen on one hand as a strength, i.e., a single head-worn sensor was sufficient to recognize the considered affective dimensions, but on the other hand, it is important to keep in mind when interpreting the head motion patterns results. Furthermore, the study is context specific, meaning that we look for aggression and negative affect in applications where we expect variations of these behaviors. Head motion can however be influenced by positive affect and by doing specific activities. Furthermore, the kind and magnitude of head motion can be person specific and also influenced by gender, culture, and personality, which we have not taken into consideration in this study. Lastly, results should be regarded while taking into consideration that we map simple head motion features to a high-level interpretation of behavior extracted from multimodal data, with varying inter-rater agreements.

7 CONCLUSION AND FUTURE WORK

This study focused on recognizing aggression and negative affect measured as levels of valence, arousal and dominance using head motion features. We demonstrated that while encapsulating both discourse and emotional information, head motion can be used to recognize affect and aggression during both listening and speaking. This is important as head motion can potentially be useful when other frequently used modalities might be unavailable or altered, e.g speech missing when listening, facial expression occluded or altered by speech. Among the tested classification approaches, the best results were achieved when taking the temporal aspect of the interactions into account, using the CRF classifier. Moreover, all classification tasks achieved higher than chance accuracies, both for the speaking and listening conditions. Aggression and Dominance were better predicted on segments where the subjects were listening, whereas Valence and Arousal were better predicted on segments where subjects were speaking.

In terms of feature importance and feature trends, the total number of head movements and RMS acceleration for pitch were among the most predictive features for all tasks. The magnitude of variation in pitch angular acceleration significantly increased with escalation of the considered dimensions for both speaking and listening. The total number of head movements significantly increased with the escalation for speaking, but significantly decreased with escalation for listening.

Head motion can be in principle easily incorporated in VR-related applications that use head-mounted displays, resulting in additional non-intrusive data collection especially required when other modalities such as facial expressions are difficult to capture. While research indicates that people use natural non-verbal behavior when interacting with realistic embodied avatars [60], the findings of this study are limited to face-to-face interactions, and further research is needed to explore behavior in VR. Other considered directions for future work include exploring head motion in combination with other modalities available on the NAA database. As overlapping speech was one of the best predictors of conflicts we are also considering joint analysis of the conversation partners.

REFERENCES

[1] A. Adams, M. Mahmoud, T. Baltrusaitis, and P. Robinson. 2015. Decoupling facial expressions and head motions in complex emotions. In *Proc. of on Affective Computing and Intelligent Interaction 2015*. IEEE Computer Society, Washington, DC, USA, 274–280.

[2] S. Alghowinem, R.d Goecke, M. Wagner, G. Parker, and M. Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *Humaine Association Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, Washington, DC, USA, 283–288.

[3] Johnie J Allen and Craig A Anderson. 2017. Aggression and violence: Definitions and distinctions. In *The Wiley handbook of violence and aggression*. John Wiley & Sons, Ltd, Chichester, UK, 1–14. https://doi.org/10.1002/9781119057574.whbva001

[4] Z. Ambadar, J.F. Cohn, and L.I. Reed. 2009. All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous. *J. of Nonverbal Behavior* 33, 1 (2009), 17–34.

[5] C. A. Anderson. 2000. Violence and aggression. In *Encyclopedia of psychology, Vol. 8.* American Psychological Association, Washington, DC, US, 162–169.

[6] Craig A. Anderson and Brad J. Bushman. 2002. Human Aggression. *Annual Review of Psychology* 53, 1 (2002), 27–51.

[7] J.-Y. Antoine, J. Villaneau, and A. Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation.. In *EACL 2014*. Association for Computational Linguistics, NY, USA, 10–p.

[8] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor. Newsl.* 6, 1 (2004), 20–29. https://doi.org/10.1145/1007730.1007735

[9] Maciej Behnke, Nadia Bianchi-Berthouze, and Lukasz D Kaczmarek. 2021. Head movement differs for positive and negative emotions in video recordings of sitting individuals. *Scientific Reports* 11, 1 (2021), 7405.

[10] Kristin P. Bennett and Colin Campbell. 2000. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explor. Newsl.* 2, 2 (2000), 1–13.

[11] Leonard Berkowitz. 1998. 3 - Affective Aggression: The Role of Stress, Pain, and Negative Affect. In *Human Aggression*, Russell G. Geen and Edward Donnerstein (Eds.). Academic Press, San Diego, 49–72.

[12] Ekaba Bisong. 2019. *Logistic Regression.* Apress, Berkeley, CA, 243–250.

[13] S.M. Boker, J.F. Cohn, B.-J. Theobald, I. Matthews, T. Brick, and J.R. Spies. 2009. Effects of damping head movement and facial expression in dyadic conversation using real–time facial expression tracking and synthesized avatars. *Philosophical Trans. of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3485–3495.

[14] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[15] Carlos Busso and Shrikanth S Narayanan. 2008. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Proc. of Ninth annual conference of the international speech communication association, Interspeech 2008.* ISCA, 1670–1673. https://doi.org/10.21437/Interspeech.2008-463

[16] A. Carroll, M. McCarthy, S. Houghton, E. Sanders O'Connor, and C. Zadow. 2018. Reactive and proactive aggression as meaningful distinctions at the variable and person level in primary school-aged children. *Aggressive Behavior* 44, 5 (2018), 431–441.

[17] M. Cristani, V. Murino, and A. Vinciarelli. 2010. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *2010 Conf. on Computer Vision and Pattern Recognition - Workshops*. IEEE Computer Society, NY, USA, 51–58. https://doi.org/10.1109/CVPRW.2010.5543179

[18] N. Cummins, J. Joshi, A. Dhall, V.n Sethu, R. Goecke, and J. Epps. 2013. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proc. of the 3rd ACM Int. workshop on Audio/visual emotion challenge*. ACM, NY, USA, 11–20.

[19] O.J. Dunn. 1964. Multiple Comparisons Using Rank Sums. *Technometrics* 6, 3 (1964), 241–252.

[20] J. O. Egede, S. Song, T. A Olugbade, C. Wang, C De C Amanda, H.g Meng, M. Aung, N. D. Lane, M. Valstar, and N. Bianchi-Berthouze. 2020. Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. In *2020 15th IEEE Int. Conf. on Automatic Face and Gesture Recognition*. IEEE Computer Society, NY, USA, 849–856.

[21] R. el Kaliouby and P. Robinson. 2005. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-Time Vision for Human-Computer Interaction*. Springer, Boston, MA, 181–200.

[22] F. Renee G. Moraga, S. Klein Tuente, S. Perrin, P. Enebrink, K. Sygel, W. Veling, and M. Wallinius. 2022. New developments in virtual reality-assisted treatment of aggression in forensic settings: The case of VRAPT. *Frontiers in Virtual Reality* 2 (2022), 174.

[23] G. Giannakakis, D. Manousos, P. Simos, and M. Tsiknakis. 2018. Head movements in context of speech during stress induction. In *Proc. of Int. Conf. on Automatic Face & Gesture Recognition*. IEEE Computer Society, NY, USA, 710–714.

[24] H. Gunes and M. Pantic. 2010. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Proc. of Intelligent Virtual Agents*. Springer, Boston, MA, 371–377.

[25] H. Gunes and M. Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *J. of Network and Computer Applications* 30, 4 (2007), 1334–1345.

[26] U. Hadar, T.J. Stiener, E.C. Grant, and F. C. Rose. 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech* 26 (1983), 117–129.

[27] U. Hadar, T.J. Stiener, E.C. Grant, and F. C. Rose. 1983. Kinematics of head movements accompanying speech during conversation. *Human Movement Science* 2 (1983), 35–45.

[28] Z. Hammal, J.F. Cohn, and D.T. George. 2014. Interpersonal Coordination of Head Motion in Distressed Couples. *IEEE Trans. on Affective Computing* 5, 2 (2014), 155–167.

[29] Z. Hammal, J.F. Cohn, C.L. Heike, and M.L. Speltz. 2015. Automatic Measurement of head and Facial Movement for analysis and Detection of infants' Positive and negative affect. *Frontier ICT* 2, 21 (2015), 11 pages.

[30] Z. Hammal, J.F. Cohn, C. Heike, and M.L. Speltz. 2015. What can head and facial movements convey about positive and negative affect?. In *2015 Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, NY, USA, 281–287.

[31] J.A. Harrigan. 1985. Listeners' Body Movements and Speaking Turns. *Communication Research* 12 (1985), 233—-250.

[32] S. Duncan Jr. 1972. Some signals and rules for taking speaking turns in conversations. *J. of Personality and Social Psychology* 23, 2 (1972), 283–292.

[33] S. Duncan Jr. and D.W. Fiske. 1977. *Face-to-face interaction: Research, methods, and theory.* L. Erlbaum Associates, New Jersey.

[34] A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63.

[35] Samuel Kim, Fabio Valente, Maurizio Filippone, and Alessandro Vinciarelli. 2014. Predicting Continuous Conflict Perceptionwith Bayesian Gaussian Processes. *IEEE Trans. on Affective Computing* 5, 2 (2014), 187–200. https://doi.org/10.1109/TAFFC.2014.2324564

[36] M.L. Knapp and J.A. Hall. 2010. *Nonverbal Behavior in Human Communication* (7 ed.). Wadsworth, Cengage Learning, Boston, MA.

[37] K. Krippendorff. 2007. Computing Krippendorff's alpha reliability. , 43 pages. Departmental papers (ASC).

[38] W.H. Kruskal and W.A. Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. of the American Statistical Association* 47, 260 (1952), 583–621.

[39] J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. 18th Int. Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 282–289.

[40] Iulia Lefter, Alice Baird, Lukas Stappen, and Björn W Schuller. 2022. A cross-corpus speech-based analysis of escalating negative interactions. *Frontiers is Computer Science, Human Media Interaction* 4 (2022), 749804.

[41] I. Lefter, G. Burghouts, and L. Rothkrantz. 2015. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Trans. on Affective Computing* 7, 2 (2015), 162–175.

[42] I. Lefter and S. Fitrianie. 2018. The Multimodal Dataset of Negative Affect and Aggression: A Validation Study. In *Proc. of the 2018 on Int. Conf. on Multimodal Interaction*. ACM, NY, USA, 376–383.

[43] Iulia Lefter and Siska Fitrianie. 2023. On Head Motion for Recognizing Aggression and Negative Affect during Speaking and Listening: Supplementary Materials. (2023). https://doi.org/10.4121/0478556b-ff3a-40ac-9662-f0e0a18ab4ee

[44] I. Lefter and C.M. Jonker. 2017. Aggression recognition using overlapping speech. In *2017 7th Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, NY, USA, 299–304.

[45] I. Lefter, C.M. Jonker, S. Klein Tuente, W. Veling, and S. Bogaerts. 2017. NAA: A multimodal database of negative affect and aggression. In *2017 7th Int. Conf. on Affective Computing and Intelligent Interaction*. IEEE Computer Society, NY, USA, 21–27.

[46] S. Madan, M. Gahalawat, T. Guha, and R. Subramanian. 2021. Head Matters: Explainable Human-centered Trait Prediction from Head Motion Dynamics. In *Proc. of the 2021 Int. Conf. on Multimodal Interaction*. ACM, NY, USA, 435–443.

[47] E. Z. McClave. 2000. Linguistic functions of head movements in the context of speech. *J. of Pragmatics* 32, 7 (2000), 855 – 878.

[48] Nathan Moore, Naseem Ahmadpour, Martin Brown, Philip Poronnik, and Jennifer Davids. 2022. Designing Virtual Reality–Based Conversational Agents to Train Clinicians in Verbal De-escalation Skills: Exploratory Usability Study. *JMIR Serious Games* 10, 3 (2022), e38669.

[49] E. Murphy-Chutorian and M.M. Trivedi. 2009. Head Pose Estimation in Computer Vision: A Survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31, 4 (2009), 607–626.

[50] S. Nayak, B. Nagesh, A. Routray, and M. Sarma. 2021. A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering* 93 (2021), 107280.

[51] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari. 2018. Survey on Emotional Body Gesture Recognition. *IEEE Trans. on Affective Computing* 12 (2018), 505–523.

[52] T. Parr, K. Turgutlu, C. Csiszar, and J. Howard. 2018. Beware Default Random Forest Importances. https://explained.ai/rf-importance/.

[53] A. Pease and B. Pease. 2017. *The definitive book of body language.* Orion, London, England.

[54] Bruno Peixoto, Bahram Lavi, João Paulo Pereira Martin, Sandra Avila, Zanoni Dias, and Anderson Rocha. 2019. Toward subjective violence detection in videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Computer Society, NY, USA, 8276–8280.

[55] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner. 2011. Anger recognition in speech using acoustic and linguistic cues. *Speech Communication* 53, 9-10 (2011), 1198–1209.

[56] P. Rani, N.n Sarkar, and J. Adams. 2007. Anxiety-based affective communication for implicit human–machine interaction. *Advanced Engineering Informatics* 21, 3 (2007), 323–334.

[57] A. Samanta and T. Guha. 2017. On the role of head motion in affective expression. In *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE Computer Society, NY, USA, 2886–2890.

[58] K. R. Scherer and T. Bänziger. 2010. On the use of actor portrayals in research on emotional expression. In *Blueprint for affective computing: A sourcebook*, K. R. Scherer, T. Bänziger, and E. B. Roesch (Eds.). Oxford university Press, Oxford, England, 166–176.

[59] R. De Silva and N. Bianchi-Berthouze. 2004. Modeling human affective postures: An information theoretic characterization of posture features. *J. Computer Animation and Virtual Worlds* 15, 3/4 (2004), 269–276.

[60] H. J. Smith and M. Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, NY, USA, 1–12.

[61] Meishu Song, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Zijiang Yang, Shuo Liu, Zhao Ren, Ziping Zhao, and Björn W Schuller. 2021. Frustration recognition from speech during game interaction using wide residual networks. *Virtual Reality & Intelligent Hardware* 3, 1 (2021), 76–86.

[62] Stephanie Tan, David MJ Tax, and Hayley Hung. 2021. Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–22.

[63] B.-J. Theobald, I. Matthews, M. Mangini, J.R. Spies, T. Brick, J.F. Cohn, , and S.M. Boker. 2009. Mapping and manipulating facial expression. *Language and Speech* 52, 2/3 (2009), 369–386.

[64] Theodoros Theodoridis and Huosheng Hu. 2013. Modeling Aggressive Behaviors With Evolutionary Taxonomers. *IEEE Trans. on Human-Machine Systems* 43, 3 (2013), 302–313. Conf. Name: IEEE Trans. on Human-Machine Systems.

[65] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, and H.C. Traue. 2018. Head movements and postures as pain behavior. *PLoS ONE* 13, 2 (2018), e0192767.

[66] T. Xue, A. E. Ali, G. Ding, and P. Cesar. 2021. Investigating the relationship between momentary emotion self-reports and head and eye movements in hmd-based 360 vr video watching. In *Extended Abstracts of the 2021 CHI Conf. on Human Factors in Computing Systems*. ACM, NY, USA, 1–8.

[67] V. Yngve. 1970. On getting a word in edgewise. *Papers from the sixth Regional Meeting of the Chicago Linguistic Society* 6-7 (1970), 567–578.

[68] Wojtek Zajdel, Johannes D Krijnders, Tjeerd Andringa, and Dariu M Gavrila. 2007. CASSANDRA: audio-video sensor fusion for aggression detection. In *2007 IEEE conference on advanced video and signal based surveillance*. IEEE Computer Society, NY, USA, 200–205.

[69] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji. 2019. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–32.