Journal on Multimodal User Interfaces (2024) 18:103–117
https://doi.org/10.1007/s12193-023-00426-5

ORIGINAL PAPER

Modelling the “transactive memory system” in multimodal multiparty interactions

Beatrice Biancardi[1] · Maurizio Mancini[2] · Brian Ravenet[3] · Giovanna Varni[4]

Received: 10 February 2023 / Accepted: 10 October 2023 / Published online: 11 November 2023

Abstract
Transactive memory system (TMS) is a team emergent state representing the knowledge of each member about “who knows what” in a team performing a joint task. We present a study to show how the three TMS dimensions Credibility, Specialisation, Coordination, can be modelled as a linear combination of the nonverbal multimodal features displayed by the team performing the joint task. Results indicate that, to some extent, the three dimensions of TMS can be expressed as a linear combination of nonverbal multimodal features. Moreover, the higher the number of modalities (audio, movement, spatial), the better the modelling. Results could be used in future work to design human-centered computing applications able to automatically estimate TMS from teams’ behavioural patterns, to provide feedback and help teams’ interactions.

Keywords Transactive memory system · Small group interactions · Human-centered computing · Nonverbal behaviour

1 Introduction

Human-centered computing (HCC) studies the human *in relation with* computing devices. It differentiates from human–computer interaction (HCI) as it deals with how the human (an individual, a team, or a society) relates to computers and other humans. Its focus is on the multi-faceted nature of humans, including emotions, social skills, attitudes, and so on. According to Clarkson et al. [21] and Canny [19], HCC is about studying “computational artifacts in support of human endeavours” and the “implications of computing in a task-directed way”, by spanning several disciplines as Computer Science and Social Sciences. One of the currently open challenges in Computer Science, specifically those related to Social Signal Processing and Affective Computing, is conceiving and building computing systems to support humans in team activities seamlessly. This process has to be informed by Social Sciences, as they have a long tradition of studying socio-affective phenomena occurring in teams.

When working together, the affective, behavioural and cognitive interaction between people often contributes to the emergence of dynamic processes called “team emergent states” [46, 55]. One of them is the transactive memory system (TMS), a cognitive team emergent state related to the specific knowledge owned by each team member. The term “transactive” highlights the relevance of exchanging information about members’ knowledge and expertise. TMS combines each member’s *personal* field of knowledge (e.g., Robert has a mathematical background, while Susan has a history in arts) with the *awareness of each other’s* one (e.g., Robert knows that Susan is specialised in arts, whereas Susan knows that Robert is good at maths) [81]. In this work, we share the conceptualisation of TMS given in [52]. That

Beatrice Biancardi and Giovanna Varni carried out the work in the paper while they were with LTCI, Télécom Paris, Institut polytechnique de Paris, Palaiseau, France.

✉ Maurizio Mancini
m.mancini@di.uniroma1.it

Beatrice Biancardi
bbiancardi@cesi.fr

Brian Ravenet
brian.ravenet@lisn.fr

Giovanna Varni
giovanna.varni@unitn.it

[1] CESI, LINEACT, 93 Boulevard de la Seine, 92000 Nanterre, France

[2] Department of Computer Science, Sapienza University of Rome, Viale Regina Elena, 295, 00161 Rome, Italy

[3] LISN-CNRS, Université Paris-Saclay, Rue du Belvédère, 91400 Orsay, France

[4] Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 9, 38123 Trento, Italy

is, TMS differentiates from other forms of socially shared cognition on different aspects. First, it depends not only on understanding "who knows what" but also on the degree to which a team's knowledge is *differentiated*. In addition, it includes the *dynamic* interplay between organised store of knowledge (TMS structure) and a set of knowledge-relevant transactive processes (encoding, storage, and retrieval processes) that occur among team's members.

It is well known that developing TMS within a team can significantly improve performance, productivity and, therefore, efficiency [3, 46, 51, 61, 63], by enabling work sharing, thus, reducing the individual cognitive load [35]. Recent findings suggest that TMS is strongly linked to affective outcomes, such as team trust, efficacy and satisfaction [84].

While there is no joint agreement on how TMS emerges within a team, all the theories about it state the relevance of interpersonal verbal and nonverbal communication.

Individuals communicating with each other are keener to select the information they are willing to learn [33] (e.g., as John is a mechanical specialist, he will be interested in car engines only). Similarly, information retrieval is facilitated if communication happens during learning [33]. This aligns with Pavitt's theory [68], which considers communication as a context for learning. Moreover, communication enables a better inter-members' understanding [14] and prevents the team from applying stereotypes about each other's expertise [36].

Although several previous studies show that nonverbal communication can predict some emergent states (e.g., cohesion) [47], to our knowledge no studies are exploring how nonverbal behaviours and TMS are related. Consequently, no work focuses on how computing systems can deal with nonverbal behaviours characterising TMS within teams. The development of such systems would envisage HCC applications to facilitate team problem-solving, and the definition of computational models for effectively supporting team collaboration.

The work presented in this paper is a first step towards this goal. Its main contribution is to investigate which nonverbal features, already exploited in studying other emergent states (e.g., leadership and cohesion, see Sect. 3), can predict TMS, both unimodally and multimodally (see Sect. 6). An overview of our approach is shown in Fig. 1. From our findings, we provide insights into the development of HCC systems leveraging TMS.

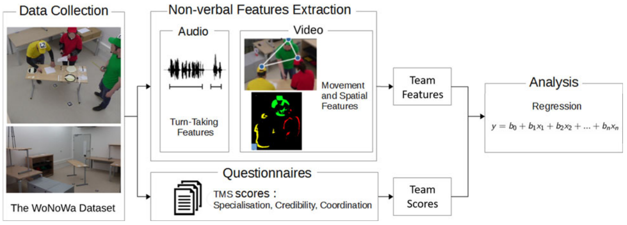
Data Collection
The WoNoWa Dataset
Non-verbal Features Extraction
Audio
Turn-Taking Features
Video
Movement and Spatial Features
Team Features
Questionnaires
TMS scores :
Specialisation, Credibility, Coordination
Team Scores
Analysis
Regression
$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$

Fig. 1 An overview of our approach. From the multimodal dataset WoNoWa (see Sect. 4), we extract nonverbal multimodal features: audio, movement and spatial arrangement of teams. The choice of these features is mainly inspired by those previously found to be relevant for estimating the team's emergent states related to TMS. We compute team features and team scores from the extracted team member's features and self-assessed TMS (see Sect. 5). Finally, we analyse the role of team features in modelling and predicting TMS by running multiple linear regression analyses (see Sect. 6)

2 Theoretical background

According to Moreland, a team includes at least 3 individuals sharing knowledge, activities and so on [62]. Unlike dyads, team interactions are more complex since they include one-to-one and one-to-many interactions. As a consequence, this complexity also applies to team "emergent states". These are defined as "cognitive, affective, and motivational states of teams that are dynamic and vary as function of team context inputs, processes, and outcomes" [55, 71]. Three categories of emergent states have been identified [30, 46]. *Cognitive* emergent states are related to the management of collective knowledge affecting team performance (e.g., Shared Mental Models [22] and Transactive Memory System). *Behavioural* emergent states are related to the activities and interactions between team members (e.g., processes related to planning, monitoring, coordination and decision-making). Finally, *Emotional* or *Affective* emergent states (e.g., cohesion and trust) include psychological states relating to feelings, attitudes, and emotions of the team members [71]. While behavioural emergent states can be directly measured from the team's behaviours, for example, by automatically extracting behaviours through sensors, and some efforts are being made towards measuring emotional emergent states from team's dynamics [73, 80], cognitive emergent states, such as TMS, have only been measured through indirect measures such as questionnaires and recall [52, 53, 64].

2.1 The transactive memory system

The transactive memory system (TMS) is an extension of an individual's memory to the team level. In other words, transactive memory refers to the awareness of one's knowledge and skills. TMS develops when each team member is also aware of the knowledge and skills of the others. So, they build a mental representation of how knowledge is distributed between each other (i.e., "who knows what"), allowing them to extend their individual knowledge [81].

TMS is a multidimensional construct consisting of: (i) *Credibility*, that is, the trust that the knowledge possessed by any of the other members is correct and accurate; (ii) Knowledge *Specialisation*, that is, the differentiation of knowledge between the team members; (iii) *Coordination*, that is, the ability of the members to work together smoothly [51, 63]. *Credibility* and *Coordination* are key factors to affective outcomes of TMS [59, 83].

The development of TMS follows the 3 phases characterising any memory system: *Encoding*, *Storage*, and *Retrieval*. During the *Encoding* phase, the team members infer "who knows what" by having multiple information exchanges. For example, a student group has to complete a project assignment. They have known each other since the 1st year and know that Robert is good at planning, Susan is very creative and Alice is good at programming. In the *Storage* phase, the team members distribute the incoming information according to each other's expertise [54]. For example, when the professor communicates the deadlines for the deliverables of the project, Robert will be particularly attentive to this information since he is the one in charge of the planning. Acceptance and shared awareness of expertise are needed in this phase. Finally, in the *Retrieval* phase, the team members know from whom they can obtain the knowledge they need [54]. For example, Alice is programming the software and asks Susan for hints for designing the user interface. Here, knowledge distribution in the team is necessary.

2.2 Interpersonal communication and TMS

Interpersonal communication in a team can be both voluntary and involuntary [27], and does not always imply verbal exchanges [60]. Previous work highlighted the role of nonverbal behaviours in team communication, including cues such as spatial arrangement, management of inter-member distances, speaking turn patterns, interruptions, etc. [23, 32, 37]. Interpersonal communication is crucial for TMS [69], being among the factors that precede [72] and support its development through all the 3 phases. In particular, some studies have shown that the use of nonverbal and para-linguistic cues in face-to-face communication allows members to signal and combine their knowledge more effectively compared to during non-face-to-face communication (e.g., computer-mediated) [34]. Communication during team training also facilitates the collective recall of information [64].

Other authors focused on the role of communication in the 3 dimensions of TMS (i.e., *Credibility*, *Specialisation*, *Coordination*). Kleanthous et al. [44] investigated how each dimension varies over time in a team navigating a 3D virtual environment collaboratively. They showed the important role of communication on *Coordination* and of gesticulation on *Credibility*.

Yoo and Kanawattanachai [82] and Rahimpour [70] noted that the amount of communication plays a crucial role in establishing TMS and, after that, its role gradually becomes less relevant. For example, to build a TMS, Yoo & Kanawattanachai asked teams to communicate remotely to manage a company's finances and thus to share different areas of expertise (marketing, finance, production, operations and human resources). They found a positive influence of communication on the development of the TMS which stopped once it was built. Argote et al. [2] highlighted that the influence of communication on TMS changes according to the presence of a team leader. They showed that teams without a leader have more robust TMS over time which leads to a better performance of the team. This result can be explained as due to the increased communication taking place in teams without a leader.

3 Related work

As mentioned above, previous work on nonverbal behaviours and emergent states that could be exploited in HCC is neglecting TMS, preferring behavioural and emotional emergent states, e.g., emergent leadership, cohesion and trust in a team. This could be explained by the fact that TMS, being a cognitive emergent state, deals with abstract knowledge (meta-memory), so it is more difficult to investigate by looking at more concrete cues (nonverbal behaviours). In this paper, we ground on the features that have already been shown to perform well in predicting other emergent states (e.g., leadership and cohesion). We hypothesise that some of them could also be related to TMS, since psychological models in the literature show the relationship between TMS and leadership [4, 48], as well as a predictive effect of task cohesion on TMS in the context of football teams [49].

Most of the works on automatic assessment of group dynamics focused on multimodal analysis of team meetings corpora, such as the ICSI [38], AMI [41], ATR [17], NTT [67] and ELEA corpus [75]. While some works focused on the prediction of individual-related dimensions such as personality [57] or individual performance [50], in the following we focus on work on the extraction of nonverbal features and their relevance in inferring behavioural and emotional emergent states. Sanchez et al. analysed the correlation between the emergence of individual leadership in team interactions (measured through questionnaires about team members' perception of each other) and acoustic [73], body/head [74] and attention [76] features. Results suggest that emergent leaders are those who talk the most, have more speaking turns and interrupt the most. They also show that body activity and motion are important in the perception of emergent leadership and that the combination of acoustic and visual information performs better than single modalities. Finally, they show that visual attention features are not better estimators of leadership than speaking activity.

More recent approaches for emergent leadership detection investigate other features and apply more complex machine learning models. Beyan et al. propose several approaches. First, they model emergent leadership by using features related to visual attention only (from head and body activity) [6]. They extract the same features used in [76] and a set of different ones, by leveraging a different and more accurate method based on head pose estimation. Then, the authors describe an approach for extracting features based on 2D pose estimation [8]. These features perform better in emerging leadership detection compared to the existing visual features. In a more recent work [9], they propose a sequential approach based on unsupervised deep learning generative models. Other works investigate the expression of different leadership styles [7, 24–26, 39].

Another emergent state is team cohesion, whose investigation in HCC is initiated by Gatica-Perez & Hung [37]. They automatically extract multimodal features (audio, visual and audio-visual) to infer low/high cohesion in task-based team meetings through machine learning techniques (e.g., SVM). Results indicate a particular relevance of turn-taking patterns. Nanninga et al. extend this work, by adding para-linguistic mimicry features and separately observing the social and task dimensions of cohesion [66]. A more recent study considers features of 3 categories: nonverbal (e.g., gaze, laughter, and so on), dialogue acts and interruptions [42]. These features are studied separately and then combined together. Results show a positive correlation with the cohesion of, among others, mutual gaze and laughter, as well as the number of speaking turns, overlaps and interruptions. More interestingly, certain behaviours that are not associated with cohesion when analysed separately, do have an impact when combined with other cues of different modalities (e.g., dialog acts with head nods). Walocha et al. explore the dynamics of task and social dimensions of cohesion, grounding on motion-capture features only [80]. They predict the decrease of cohesion over time, by using self-reported annotations of team cohesion as labels. Their results highlight a (positive or negative) impact of the maximum distance between team members, the overall posture expansion and the amount of facing between each person. In addition, some features are found to be correlated to both task and social cohesion.

To summarise, the works described above show the effectiveness of using nonverbal behaviour in addressing emergent states. In particular, multimodal approaches seem to generally perform better than unimodal ones.

4 The WoNoWa dataset

WoNoWa (*W*ho k*No*ws *W*h*a*t) is a multimodal (audio and video) dataset of interactions within 15 teams, performing several activities [11]. The dataset includes automatically extracted features and manual annotations of team members' nonverbal behaviours, as well as self-assessment measures of TMS.

WoNoWa was designed to address the 3 phases of TMS, i.e., *Encoding*, *Storage* and *Retrieval* (see Sect. 2). In the *Encoding* phase, the team was given a list with 3 fields of expertise and each member could choose the preferred one. These fields were: Logistical, Mathematical and Manual expertise. In the *Storage* phase, each team member watched a brief tutorial about the chosen field of expertise.

We focus, here, on the interactions related to the *Retrieval* phase. During this phase, the team members were together in the *Interaction Area* shown in Fig. 2. The *Retrieval* phase consisted of three steps, after which each participant filled out a TMS questionnaire (see Sect. 4.2). At the beginning of this phase, each team member was asked to accomplish a task related to the chosen field of expertise: setting up the table by following the rules described in the tutorial (Logistical expertise); computing conversions between the Imperial and the International System (Mathematical expertise); making origami (Manual expertise). Then, as a team, they were asked to modify the setup of the table and to do new origami, this time following a list of dimensions (given in the Imperial system). The participants were only provided with measuring tools in the International System (meters), thus mathematical expertise was needed to accomplish the task.

Finally, the last step of the *Retrieval* phase, the *step on which we focus in this work*, they were free to self-assign the same 3 tasks in any way they wanted (but they could not choose the one they just performed). So, the members needed each other's expertise, resulting in collaboration and interaction between them. We focus on this step of the Retrieval phase, hereinafter called "interaction", for two main reasons: (1) it is the last one, so the team had the time to develop TMS through the previous ones (as confirmed by the high score of *Specialisation* and *Coordination*, as well as higher scores of *Credibility* compared to the previous steps of the *Retrieval* phase [11]); (2) the team members are engaged in a collaborative task requiring a high level of interaction.

4.1 Technical setup

WoNoWa was collected in an experimental room depicted in Fig. 2. A table was placed in the center of the *Interaction Area*, while two more tables were placed in the corners of the room. Team members performed the tasks related to the different fields of expertise as indicated in Fig. 2. The team interaction was recorded via 3 video cameras at 1920 × 1080, progressive scan, 50 fps. Two of them were installed at a height of 3 ms in the opposite corners of the room, so each member could be viewed by at least one camera at all times. However, each video camera could capture only a part of the room, so camera view fusion had to be performed, as described in Sect. 5.2.1. An additional frontal video camera

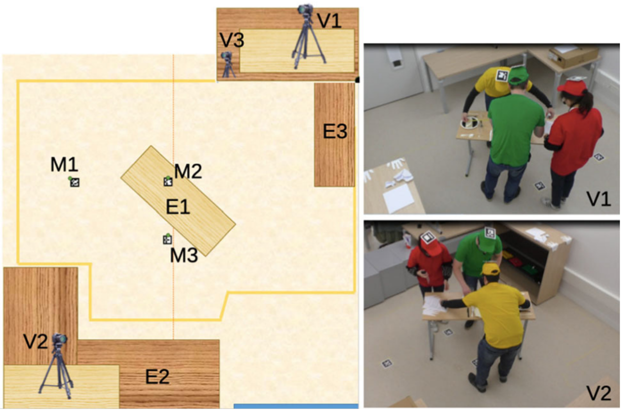

Fig. 2 On the left: plan of the *Interaction Area*, measuring 3.90 by 3.87 ms. On the top right: view from the video camera placed in the North-East corner of the area (V1). On the bottom right: view from the video camera placed in the South-West corner of the area (V2). In each view, the ArUco Markers [29] (M1, M2, M3) can be viewed by the corresponding camera. Each table corresponds to one of the 3 fields of expertise: Logistical (E1), Mathematical (E2) and Manual (E3)

was positioned at a lower height to provide an additional global view of the area, to facilitate the manual annotations (see Sect. 5.3). The video cameras were calibrated to correct the white balance and compensate for the lens distortion.

For tracking the team members' positions in the room, we used ArUco markers, fiducial markers based on a seven-by-seven binary grid [29]. Three reference markers were positioned on the floor in a way that they were visible by both cameras and remained constant throughout the experiment (see M1, M2 and M3 in Fig. 2). One unique ArUco marker was placed on each of the baseball hats worn by each team member (see Sect. 5.2.1). Each participant wore a wireless microphone headset recording at 44.1k Hz, in separate channels. They also wore t-shirts of different colours to facilitate the extraction of upper body silouhette used to compute movement features (see Sect. 5.2.2).

4.2 Self-assessment scores of TMS

The team members were asked to fill out a questionnaire about their perception of TMS in the team, after each step of the *Retrieval* phase. The questionnaire contained Lewis' items [51] measuring the 3 dimensions of TMS (i.e., *Credibility*, *Specialisation*, *Coordination*). For French participants, the French translation of Lewis' questionnaire, validated by Michinov [58], was used. All the scores were given on a 5-point Likert scale, where 1 stands for "I totally disagree" and 5 stands for "I totally agree".

For each TMS dimension, Cronbach α was computed to measure the reliability of the items. Two items from the *Coordination* sub-scale were discarded since they were negatively correlated with the others belonging to the same sub-scale, indicating that the team members did not rightly interpret them. The α computed on the remaining items indicated *acceptable* to *very good* reliability (0.83 for *Credibility*, 0.78 for *Specialisation* and 0.67 for *Coordination*). The scores of each item of the same sub-scale were then averaged to obtain one score per member.

To obtain one score per team and per TMS dimension (i.e., one score for *Credibility*, one for *Specialisation*, and one for *Coordination*), we checked whether the team members agreed about the score of TMS dimension they assigned to their team. ICCs (two-way, average) with consistency definition [45] were computed for each team, revealing a *fair* to *excellent* agreement (fair for 2 teams, good for 2 teams and excellent for 10 teams, all $p < 0.001$) [20] except for one team (ICC $= -0.66$, $p = 0.97$), that was excluded from the analyses. Finally, for each TMS dimension and each team, we computed the mean of the team member scores.

5 Nonverbal features extraction

As mentioned in Sect. 4, we focus on nonverbal features extracted from data collected in the last step of the *Retrieval* phase.

In the remainder of this Section, we describe the nonverbal features organised according to the modality they belong to: Audio, Movement and Spatial. Table 1 summarises their descriptive statistics.

5.1 Audio features

Audio features were extracted from the audio recordings and are related to vocal turn-taking, which plays an important role in developing social dimensions like competition and collaboration [32, 40]. Vocal turn-taking includes silence, silence overlaps, and interruptions. In particular, interruptions are a relevant cue in face-to-face conversations: they can be considered as turn-taking violations [5], reflecting interpersonal attitudes (e.g., dominance or cooperation) as well as engagement in the interaction [65].

5.1.1 Pre-processing

To compute the audio features relative to team members' turn-taking activity, we applied a series of transformations to the raw audio files. The audio recordings of each team member were manually synchronised with the videos by referring to a clap that the experimenter performed at the beginning and at the end of each recording. The raw files were normalised/compressed and a noise reduction filter was applied in Audacity.[1] Additionally, about 10% of the files

[1] https://www.audacityteam.org.

Table 1 Descriptive statistics of the features, organised in 3 categories: Audio (A), Movement (M), Spatial (S)

Name	Feature	Mean	SD	Min	Max
Audio					
Total speaking turns [turn/min]	*TST*	6.65	0.84	5.04	7.76
Total speaking length [s/min]	*TSL*	18.99	3.36	0.12	25.31
Average speaking turn [s/turn]	*AST*	2.96	0.57	1.90	3.92
Total attempted interruptions	*TAI*	3.54	0.99	2.07	5.62
Total successful interruptions	*TSI*	1.69	0.56	0.88	3.21
Successful interruptions percentage	*SIP*	49.99	6.75	3.60	61.67
Movement					
Head velocity mean [px/f]	*HV*	0.17	0.06	0.11	0.32
Head velocity standard deviation [px/f]	*HVSD*	0.13	0.05	0.09	0.27
Head directness mean	*HDir*	0.18	0.06	0.09	0.35
Head directness standard deviation	*HDirSD*	0.13	0.07	0.07	0.35
Quantity of motion mean [px^2/f]	*QoM*	0.03	0.01	0.02	0.05
Quantity of motion standard deviation [px^2/f]	*QoMSD*	0.02	0.00	0.01	0.02
Head velocity entropy mean	*HVE*	0.01	0.00	0.00	0.01
Head velocity entropy standard deviation	*HVESD*	0.00	0.01	0.00	0.01
Quantity of motion entropy mean	*QoME*	0.00	0.00	0.00	0.01
Quantity of motion entropy standard deviation	*QoMESD*	0.00	0.00	0.00	0.00
Head distance mean [px/f]	*HDist*	1.24	0.29	0.60	1.71
Head distance standard deviation [px/f]	*HDistSD*	1.83	0.22	1.32	2.06
Spatial					
Semi-circular f-formations frequency [evt/min]	*SCffF*	2.55	1.00	1.00	5.00
Semi-circular f-formations mean time [s/evt]	*SCffT*	23.58	15.81	1.38	67.33
Semi-circular f-formations percentage	*SCffP*	12.48	9.08	0.52	34.22
Triangular f-formations frequency [evt/min]	*TrffF*	2.76	1.19	1.00	5.00
Triangular f-formations mean time [s/evt]	*TrffT*	24.50	23.20	6.00	78.00
Triangular f-formations percentage	*TrffP*	14.54	14.76	1.12	50.79
Other f-formations frequency [evt/min]	*OthffF*	0.21	0.43	0.00	1.00
Other f-formations mean time [s/evt]	*OthffT*	0.78	1.76	0.00	6.00
Other f-formations percentage	*OthffP*	0.13	0.29	0.00	0.97
Personal area occupation frequency [evt/min]	*PAF*	2.71	1.31	1.00	5.00
Personal area occupation mean time [s/evt]	*PAT*	37.74	27.95	2.92	107.31
Personal area occupation percentage	*PAP*	24.25	16.78	0.57	58.34
Others’ area occupation frequency [evt/min]	*OAF*	2.50	1.16	0.33	4.67
Others’ area occupation mean time [s/evt]	*OAT*	17.50	12.08	0.34	41.93
Others’ area occupation Percentage	*OAP*	19.64	18.21	0.66	71.89
Common area occupation frequency [evt/min]	*CAF*	3.73	1.20	1.67	5.67
Common area occupation mean time [s/evt]	*CAT*	39.66	18.49	8.33	72.83
Common area occupation percentage	*CAP*	33.53	12.43	11.51	58.44

The non-SI (International System of Units) units are: *turn* number of speaking turns, *px* pixels, *f* frame, *evt* occurrences. The features are computed over the whole video

were processed to reduce specific noises like, e.g., breathing, electromagnetic interference, and so on. To detect speaking activity, the Silence Finder function of Audacity was applied to automatically detect and mark segments exceeding a defined sound threshold. The segments were manually checked and tuned to ignore irrelevant sounds, e.g., impacts with the microphone, objects falling on the ground and non-verbal vocal behaviour (sighing, laughing, self-talking, etc). The resulting segments were binarised, with 1 representing speech and 0 non-speech.

5.1.2 Output

From the binary segmentation, we computed the following features per team member over the whole video, taking inspiration from previous work on team's analysis [37, 74]:

- *Total Speaking Turns* (number of turns per minute, per member m) - TST_m: the number of speaking turns, normalised by the interaction length in minutes;
- *Total Speaking Length* (number of seconds per minute, per member m) - TSL_m: the total speaking time, in seconds, divided by the interaction length in minutes;
- *Average Speaking Turn* (per member m) - AST_m: the average speaking turn duration, in seconds, with $AST_m = TSL_m/TST_m$;
- *Total Attempted Interruptions* (per minute, per member m) - TAI_m: the number of attempted interruptions, normalised by the interaction length. An attempted interruption occurred if a team member started speaking while another one was already speaking, resulting in an overlap;
- *Total Successful Interruptions* (per minute, per member m) - TSI_m: the number of successful interruptions, normalised by the interaction length. A successful interruption occurred if (1) a team member started speaking while another one was already speaking, resulting in overlap, and, consequently, (2) that team member stopped speaking before ending their turn;
- *Successful Interruptions Percentage* (per member m) - SIP_m: the percentage of successful over attempted interruptions, with $SIP_m = TSI_m/TAI_m * 100$.

The above features, computed on each team member, were then averaged to obtain the following team audio features: *TST*, *TSL*, *AST*, *TAI*, *TSI*, *SIP*.

5.2 Movement features

The following Movement features are computed: Head Velocity (*HV*), Head Distance (*HDist*), Head Directness (*HDir*), Entropy of HV (*HVE*), Quantity of Motion (*QoM*), and Entropy of QoM (*QoME*). The selection of these features is inspired from previous work on social interaction [16, 28, 79]

5.2.1 Head position features

Team members' head position and rotation were tracked through a marker-based approach. Each team member wore a cap with an ArUco marker [29] attached on the top. Three additional reference markers were positioned on the floor in a way that they were visible by both cameras and remained constant throughout the experiment (see M1, M2 and M3 in Fig. 2). These markers were used as references to compute the position of the members' head markers in the room. Since each camera performed a separate head tracking, they had to be merged before using them. The processing was carried out via a Python script using the OpenCV library [15]. We applied linear interpolation and average smoothing in case of missing frames. For each video frame and team member, the following data were extracted: the 3D head position (meters) $HP = (HP_x, HP_y, HP_z)$ and the 3D head rotation (radians) $HR = (HR_x, HR_y, HR_z)$.

Head velocity (HV) We computed Head Velocity HV as the magnitude of the 1st derivative of the head position:

$$HV = \sqrt{\left(\frac{dHP_x}{dt}\right)^2 + \left(\frac{dHP_y}{dt}\right)^2 + \left(\frac{dHP_z}{dt}\right)^2} \tag{1}$$

To reduce noise, we applied a Savitzky–Golay low-pass filter (order 1, frame size 75).

Head distance (HDist) For each team member, we averaged the Euclidean distance of their HP with each of the other 2 team members' HP, obtaining $HDist_i$, $HDist_j$ and $HDist_k$ for each team member, respectively. We then averaged HP_i, HP_j and HP_k for each team to obtain the team feature *Hdist* for each frame.

Head directness (HDir) The Directness of movement is a feature that estimates how much direct vs indirect a trajectory is [1, 16]. We computed Head Directness on HP trajectory over time 15 s long moving windows, with 3 s overlap:

$$HDir = \frac{||HP_{W-1} - HP_0||}{\sum_{f=0}^{W-1} ||HP_{f+1} - HP_f||} \tag{2}$$

where W is the window length (in frames) and $||\,||$ is the Euclidean distance between the head position HP in two generic frames of the window. So, $HDir$ tends to 1 if the length of the head trajectory in the time window tends to be equal to the distance between the position of the head in the first and the last frame of the window (i.e., the head trajectory is direct); it tends to 0 if the length of the head trajectory is greater than the distance between the position of the head in the first and the last frame of the window (i.e., the head trajectory is indirect).

5.2.2 Silhouette blob features

To exploit colour thresholding to detect upper body (head, torso, and arm movement) movement features, the team members wore coloured t-shirts and baseball hats. For each video frame, the upper body Silhouette Blob (SB) was extracted as the binary threshold of the HSV pixel data, and a median filter was applied to remove noise [18].

Quantity of motion (QoM) Quantity of Motion (QoM) is a 2D measure of the amount of performed movement [18]. First,

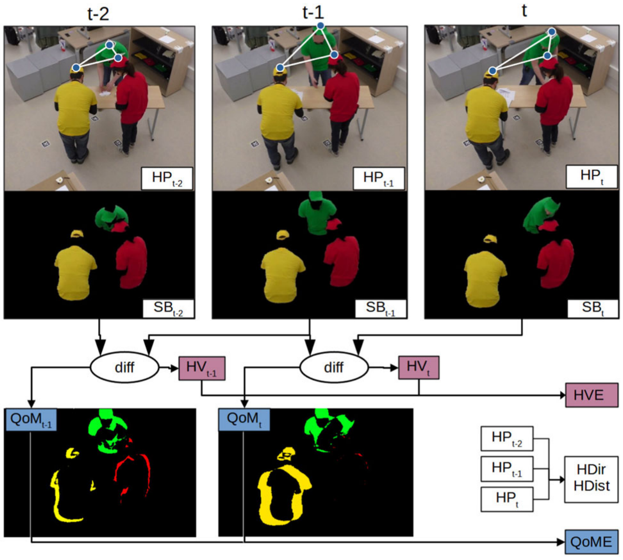

Fig. 3 The movement feature extraction framework. From top to bottom, with time flowing from left to right: video frames are read, Head Positions (HP) and Silhouette Blobs (SB) are extracted, movement features (HV, QoM, HVE, QoME, HDir and HDist) are computed

we computed the area of the binary image resulting from the XOR between 2 consecutive SBs (XOR image area). Then, *QoM* is equal to the ratio between the XOR image area and the area of the binary image resulting from the OR between the same 2 consecutive SBs. So, *QoM* tends to 0 if team members are still, it is greater than 0 if they are moving (the upper limit being 1).

5.2.3 Head velocity and quantity of motion entropy

As detailed in [79], Sample Entropy (SampEn) is a non-linear entropy extraction technique that was developed to quantify behaviour regularity by taking into account the "recent" movement history. Higher values of SampEn are associated with the higher disorder, while smaller values indicate regularity.

We used SampEn to estimate the degree of regularity of a team member's *HV* and *QoM*, that we consider as an approximation of team coordination (one of the components of TMS, see Sect. 2). We adopted the SampEn Matlab implementation described in [56] with parameters: Embedding Dimension $m = 3$, Tolerance $r = 0.2$.

SampEn was computed on moving time windows of 15 s with 3 s overlap.

Figure 3 illustrates the movement feature extraction by providing a high-level representation of the process, and by highlighting the main data and features computed at each step. From top to bottom, with time flowing from left to right: video frames are read, and Head Positions (*HP*) and Silhouette Blobs (*SB*) are extracted.

5.2.4 Output

The resulting movement features per team member m computed over the whole video are:

- Head Velocity Mean (HV_m) and Standard deviation ($HVSD_m$);
- Head Directness Mean ($HDir_m$) and Standard Deviation ($HDirSD_m$);
- Quantity of Motion Mean (QoM_m) and Standard deviation ($QoMSD_m$);
- Head Velocity Entropy Mean (HVE_m) and Standard Deviation ($HVESD_m$);
- Quantity of Motion Entropy Mean ($QoME_m$) and Standard Deviation ($QoMSD_m$).

Similarly to the audio ones, the above features computed on each team member were averaged to obtain the following team movement features: *HV*, *HVSD*, *HDir*, *HDirSD*, *QoM*, *QoMSD*, *HVE*, *HVESD*, *QoME*, *QoMSD*.

Additionally, Head Distance Mean (*HDist*) and Standard Deviation (*HDistSD*) were also computed over the whole interaction.

5.3 Spatial features

People's arrangement in the physical space (also called, F-formation) can reflect their roles in the team and the ongoing interaction [23, 43]. Studies also show that interpersonal distance changes according to the degree of closeness among people [31]. For this reason, WoNoWa includes manually annotated features (performed by 2 raters, more details in [11]) related to the team arrangement in the experimental room and to how the members occupy the different areas of the room while performing the experiment tasks.

5.3.1 F-formations

The most frequent F-formations emerging from a visual analysis were the Semi-circular and the Triangular ones. The least frequent arrangements, that is, the L-shape or the Side-by-side one, were merged into a category called Other. An example of each F-formation is shown in Fig. 4. Two identical F-formations occurring in less than 5 s were considered uninterrupted.

5.3.2 Task-related area occupation

We considered 3 main categories of Task-related Area Occupation, for each team member: when the member worked on their task (Personal Area) when the member was in the area related to a different task (Others Area), and when the member did common tasks not related to particular expertise, such

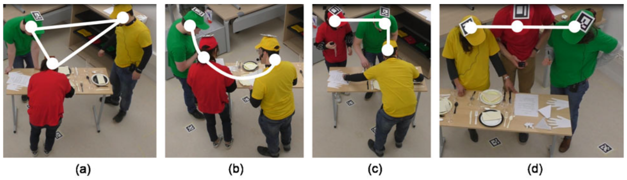

Fig. 4 An example of the F-formations annotated in the WoNoWa dataset: **a** Triangular, **b** Semi-circular, **c** L-shape, **d** Side-by-side. Since their frequency was low, **c, d** were merged into a category called Other

as reading instructions, checking the table, thinking, and so on (Common Area).

5.3.3 Output

For each F-formation and Task-related Area Occupation category, we computed its frequency (i.e., the number of occurrences per minute), the mean time duration of each occurrence (computed as the sum of all the time spent in that category divided by the total number of occurrences of that category) and the percentage of the time in which the team was engaged in that category during the task (computed as the sum of all the time spent in that category divided by the overall length of the interaction), over the whole video.

The resulting final features were:

- Semi-circular F-formations Frequency (*SCffF*), Mean Time (*SCffT*) and Percentage (*SCffP*);
- Triangular F-formations Frequency (*TrffF*), Mean Time (*TrffT*) and Percentage (*TrffP*);
- Other F-formations Frequency (*OthffF*), Mean Time (*OthffT*) and Percentage (*OthffP*);
- Personal Area Occupation Frequency (*PAF*), Mean Time (*PAT*) and Percentage (*PAP*);
- Others Area Occupation Frequency (*OAF*), Mean Time (*OAT*) and Percentage (*OAP*);
- Common Area Occupation Frequency (*CAF*), Mean Time (*CAT*) and Percentage (*CAP*).

6 Analyses and results

This work aims to model the three TMS dimensions as a linear combination of nonverbal multimodal features. We also seek to obtain explainable models with meaningful variables' interpretation (that is, that could also be meaningful for a human observer). We adopt a multiple regression analysis since this method enables identifying significant relationships between a dependent variable (TMS dimensions) and independent variables (nonverbal features). In addition, this method enables computing the strength of the impact of multiple independent variables on the dependent variable.

First, we check whether the data meets the assumptions for multiple linear regression. We remove the features causing multi-collinearity issues, that is, those that are highly correlated with each other ($r > 0.8$). This features are: *TAI*; *OAP*; *OthffP*; *OAT*; *TrffP*; *SCffP*; *HVSD*; *HDirSD*. Then, for each dependent variable (i.e., TMS dimension), we remove the features violating the linearity assumption, as follows:

- Dimension 1—Credibility: *TSL*, *OthffF*, *SCffF*, *QoM*, *QoME*;
- Dimension 2—Specialisation: none;
- Dimension 3—Coordination: *TSL*, *OthffF*, *OthffT*, *SCffF*.

Finally, we perform regression diagnostics to check for the normality and the homoscedasticity of the residuals, by running a Shapiro–Wilk and a Breusch–Pagan test, respectively. For all the results presented in this Section, the 2 latter assumptions are met (all $p > 0.05$ for Shapiro–Wilk and Breusch–Pagan tests; all correlations between observed residuals and expected residuals under normality ≥ 0.9). Since the data fit the assumptions, we use multiple linear regression.

As only a single TMS self-assessment score is given by each member at the end of each task, an issue arises as we do not have continuous assessment scores. Moreover, the number of features is significantly higher than the number of teams, and, consequently, the given assessment scores. Thus, we follow a stepwise approach to select the best predictors for each target variable.

We perform the regression models with 1, 2, or 3 modalities (i.e., Audio, Movement, or Spatial only, 2 of them or all the 3 modalities together) and, at most, 4 features (due to the small number of teams compared to the number of features). Then, from all the significant regression models (i.e., those having a p-value < 0.05 for every feature), we identify the best ones (i.e., those having the highest R^2 score) for each number of features. We then checked the predicting performance of these models by running a 10-run 5-fold cross validation. These values were chosen according to previous work [10].

Table 2 reports the selected regression models for each dimension: *Credibility*, lines 1–8; *Specialisation*, lines 9–17 and *Coordination* 18–22.

7 Discussion

For each modality and TMS dimension, we discuss here the regression models with the highest R^2. Since the feature values vary in different ranges, the β values cannot be directly compared among them. So, the reported β values only provide the direction of the correlation with the corresponding TMS score.

Table 2 The significant regression models for each TMS dimension, according to the number of features and modality: Audio (A), Spatial (S), Movement (M)

Model	Modalities	# Features	Significant features	R^2	p-value	RMSE
Dimension 1: Credibility						
1	A	1	$TST\ast\ast$	0.44	0.005	0.29
2	A+S	2	$TST\ast\ast\ast$; $OAF\ast\ast$	0.70	< 0.001	0.22
3	A+S	3	$TST\ast\ast\ast$; $OAF\ast\ast$; $PAF\ast$	0.80	< 0.001	0.17
4	S+M	2	$HV\ast$; $CAF\ast\ast$	0.50	0.009	0.27
5	S+M	3	$HV\ast$; $QoMESD\ast$; $SCffT\ast\ast$	0.49	0.020	0.34
6	S+M	4	$HV\ast$; $HDist\ast$; $QoMESD\ast$; $SCffT\ast\ast\ast$	0.66	0.007	0.29
7	A+S+M	3	$TST\ast\ast\ast$; $OAF\ast\ast\ast$; $HDistSD\ast$	0.81	< 0.001	0.19
8	A+S+M	4	$TST\ast\ast\ast$; $OAF\ast\ast\ast$; $PAF\ast\ast\ast$; $HVE\ast\ast$	0.93	< 0.001	0.11
Dimension 2: Specialisation						
9	S	2	$PAP\ast$; $TrffT^{\dagger}$	0.47	0.029	0.67
10	S	3	$CAF\ast$; $OA\ast$; $TrffT\ast$	0.57	0.031	0.46
11	A+S	2	$AST\ast\ast$; $OAT\ast\ast$	0.61	0.006	0.3
12	A+S	3	$AST\ast\ast$; $CAP\ast$; $OAT\ast\ast$	0.78	0.001	0.25
13	A+S	4	$AST\ast\ast\ast$; $CAP\ast\ast$; $OAT\ast\ast\ast$; $PAF\ast$	0.88	< 0.001	0.18
14	S+M	3	$HDist\ast$; $CAF\ast$; $OthffT\ast$	0.62	0.012	0.4
15	S+M	4	$HVESD\ast\ast$; $HDir\ast$; $OAF\ast$; $PAF\ast$	0.67	0.028	0.36
16	A+S+M	3	$AST\ast\ast\ast$; $OAT\ast\ast$; $HDir\ast$	0.76	0.002	0.3
17	A+S+M	4	$AST\ast\ast\ast$; $OAT\ast\ast\ast$; $SCffT\ast\ast$; $HDist\ast\ast$	0.89	< 0.001	0.22
Dimension 3: Coordination						
18	A	1	$AST\ast$	0.40	0.015	0.5
19	A+M	2	$TST\ast\ast$; $HDist\ast$	0.59	0.008	0.48
20	A+M	3	$AST\ast\ast$; $TSI\ast$; $HDist\ast$	0.69	0.007	0.47
21	S+M	2	$HDist\ast$; $CAF\ast$	0.53	0.016	0.56
22	A+S+M	3	$TST\ast$; $CAF\ast$; $HDist\ast\ast$	0.73	0.004	0.42

The significant features, R-squared and p-value of each model are reported. † stands for $p = 0.05$; *stands for $p < 0.05$; **stands for $p < 0.01$; ***stands for $p < 0.001$. The last column reports the RMSE for the 10-run 5-fold cross-validation

Let us consider a feature F_1 that varies in [0, 1000] and has a $\beta = 0.001$: so, F_1 has a contribution of 0.001 on the dependent variable (one of TMS dimensions), that is, on average, it contributes for $500 * 0.001 = 0.5$. Let us now consider another variable F_2 that varies in [0, 1] and has a $\beta = 1$: so, F_2 has a contribution of 1 on the dependent variable, that is, on average, it contributes for $0.5 * 1 = 0.5$. So, the two variables, despite having highly different β values (0.001 vs 1), cause the same amount of change on the dependent variable (0.5).

To quantify the impact of each feature on the TMS score dimensions, and thus enable comparison, we also report a coefficient I, representing the contribution of the features on the dependent variable, computed by multiplying β by the mean of the feature. β and I coefficients are reported in Table 3.

7.1 Dimension 1: credibility

Credibility was defined as the trust that the knowledge possessed by any of the other members is correct and accurate (see Sect. 2.1). Audio is the only modality that allows for obtaining significant unimodal models of this dimension. The feature that best models *Credibility* (model 1) is *TST* (total speaking turns). The feature is negatively correlated with *Credibility* ($\beta = -0.31$, $I = -2.06$), which could indicate that high *Credibility* implies that participants need less to ask and discuss the task among them, so they trust each other.

When looking at 2 modalities together, the best models include Spatial separately combined with Audio and Movement.

Concerning Audio and Spatial, with 2 features (model 2), *TST* is significant ($\beta = -0.33$, $I = -2.19$), as well as *OAF* (other's area occupation frequency), that is positively correlated with *Credibility* ($\beta = 0.16$, $I = 0.40$). This result

Table 3 β and I values for the regression models with the highest R^2 discussed in Sect. 7

Model	Features	Beta	I
Dimension 1: Credibility			
1	*TST*	– 0.31	– 2.06
2	*TST*	– 0.33	– 2.19
	OAF	0.16	0.40
3	*TST*	– 0.36	– 2.34
	PAF	– 0.09	– 0.24
6	*HV*	– 4.37	– 0.74
	HDist	– 0.56	– 0.69
	QoMESD	455.08	1.36
	SCffT	– 0.02	– 0.47
8	*TST*	– 0.33	– 2.19
	OAF	0.17	0.43
	PAF	– 0.13	– 0.35
	HVE	157.28	0.08
Dimension 2: Specialisation			
9	*PAP*	0.0004	0.003
	TrffT	0.008	0.96
13	*AST*	0.32	0.94
	CAP	0.015	0.5
	OthffT	0.015	0.07
	PAF	– 0.13	– 0.35
15	*HVESD*	445.29	1.33
	HDir	– 6.76	– 1.21
	OAF	0.26	0.65
	PAF	– 0.21	– 0.5
17	*AST*	0.57	1.68
	OAT	– 0.03	– 0.08
	SCffT	– 0.015	– 0.35
	HDist	– 0.67	– 0.83
Dimension 3: Coordination			
18	*AST*	0.72	2.13
19	*TST*	– 0.48	– 3.19
	HDist	– 1.19	– 1.48
20	*AST*	0.71	2.1
	TSI	– 0.52	– 0.88
	HDist	– 1.13	– 1.4
21	*HDist*	– 1.34	– 1.66
	CAF	0.32	1.19
22	*TST*	– 0.37	– 2.46
	CAF	0.22	0.82
	HDist	– 1.37	– 1.70

could be explained by the presence of helping behaviour and the team members trusting each other's expertise: one member who needs help enters the area of the person with the needed expertise.

Considering 3 features (model 3), *TST* and *OAF* are still significant ($\beta = -0.36, I = -2.34$ and $\beta = 0.17, I = 0.43$, respectively) and *Credibility* is also negatively correlated with *PAF* (personal area occupancy frequency), $\beta = -0.09, I = -0.24$). In line with what we observed with the 2-feature model, it could mean that people working alone do not seek the help of the other team members.

Moving to Spatial and Movement, the best model is the one with 4 features (model 6): *HV* ($\beta = -4.37, I = -0.74$), *HDist* ($\beta = -0.56, I = -0.69$), *QoMESD* ($\beta = 455.08, I = 1.36$) and *SCffT* ($\beta = -0.02, I = -0.47$). All these features are negatively correlated with *Credibility*. That is, team members tend to interact in a calm and steady manner.

The best model for *Credibility* is obtained when combining the 3 modalities together and considering 4 features (model 8). In particular, the significant features are: *TST* ($\beta = -0.33, I = -2.19$), *OAF* ($\beta = 0.17, I = 0.43$), *PAF* ($\beta = -0.13, I = -0.35$) and *HVE* (head velocity entropy mean, $\beta = 157.28, I = 0.08$). This result is similar to one of the models involving the 2 modalities described above.

To summarise, results show that we can estimate Credibility by looking at how much the team members communicate with each other in a confident way, which results in a low number of speaking turns and movements.

7.2 Dimension 2: specialisation

Specialisation was defined as the differentiation of knowledge between the team members (see Sect. 2.1). Spatial is the only modality that allows for obtaining significant unimodal models of this dimension. The features that best model *Specialisation* (model 9) are *PAP* (personal area occupation percentage) and *TrffT* (triangular f-formation mean time), which are both positively correlated with *Specialisation* ($\beta = 0.0004, I = 0.003$ and $\beta = 0.008, I = 0.96$, respectively). So, the impact of *PAP* is very low, while a higher impact of *TrffT* indicates that in a specialised team, the members tend to arrange in the space by following triangular configurations.

The best model using Audio and Spatial features (model 13) includes *AST* (average speaking turn, $\beta = 0.32, I = 0.94$), *CAP* (common area occupation percentage, $\beta = 0.015, I = 0.5$), *OthffT* (other f-formation mean time, $\beta = 0.015, I = 0.07$) and *PAF* (personal area occupation frequency, $\beta = -0.13, I = -0.35$). This result could mean that when *Specialisation* is high, the team members spend more time together in the common area of the room, engaging in longer speaking turns (e.g., for explaining tasks).

The best model using Movement and Spatial features (model 15) includes *HVESD* (head velocity entropy standard deviation, $\beta = 445.29, I = 1.33$), *HDir* (head directness mean, $\beta = -6.76, I = -1.21$), *OAF* ($\beta = 0.26, I = 0.65$) and *PAF* ($\beta = -0.2I = -0.5$). The positive correlation between *Specialisation* and *HVESD*, and the negative one between the same dimension and *HDir*, might indicate that team members' movements constantly vary following non-linear trajectories. These results could mean that members with high expertise go and help other members in their area of expertise.

Considering 3 modalities, the best model (model 17) includes: *AST* ($\beta = 0.57, I = 1.68$), *OAT* (other's area occupation mean time, $\beta = -0.03, I = -0.08$), *SCffT* (semi-circular f-formation mean time, $\beta = -0.015, I = -0.35$) and *HDist* (head distance mean, $\beta = -0.67, I = -0.83$). This result is complementary with the findings about *Credibility*, indicating that people go to others' areas to share their expertise.

Results show that movement features of the team members across the different areas can be used to estimate Specialisation. In particular, for high values of Specialisation, the team members' movements continuously vary following non-linear trajectories.

7.3 Dimension 3: coordination

Coordination was defined as the ability of the members to work together smoothly (see Sect. 2.1). Audio is the only modality that allows for obtaining significant unimodal models of *Coordination*. The Audio feature that best models *Coordination* (model 18) is *AST* ($\beta = 0.72, I = 2.13$). So, in a highly coordinated team, speaking turns last longer.

Moving to Audio and Movement, we obtain 2 significant models having 2 or 3 features, respectively. The former (model 19) includes *TST* ($\beta = -0.48, I = -3.19$) and *HDist* ($\beta = -1.19, I = -1.48$); the latter (model 20) includes *AST* ($\beta = 0.71, I = 2.1$), *TSI* (total successful interruptions, $\beta = -0.52, I = -0, 88$) and, again, *HDist* ($\beta = -1.13, I = -1.4$). Results show that high *Coordination* corresponds to a small number of speaking turns between the team members and a decreased distance between them.

Another significant model with 2 modalities (model 21), combines Movement with Spatial features. In particular, it includes *HDist* ($\beta = -1.34, I = -1.66$) and *CAF* (common area occupation frequency, $\beta = 0.32, I = 1.19$), similarly to what we obtained by combining Audio and Movement.

The best model for *Coordination* is obtained by combining 3 modalities (model 22). In line with the results obtained by considering models with 2 modalities, the significant features are *TST* ($\beta = -0.37, I = -2.46$), *CAF* ($\beta = 0.22, I = 0.82$) and *HDist* ($\beta = -1.37, I = -1.70$).

On the whole, results show that, in a highly coordinated team, members engage in fewer and longer speaking turns, with few interruptions. Moreover, the members tend to stay close to each other and perform activities related to the coordination of the tasks (by working in the common area).

8 Conclusion and perspectives

This paper provides the first insights on how to automatically model the three dimensions (*Credibility*, *Specialisation* and *Coordination*) of TMS as a linear regression of nonverbal features of small teams. More specifically, we focus on features of 3 modalities: audio, movement and spatial arrangement.

Moreover, linear regression is chosen to obtain explainable models. Therefore, we focus on achieving high-perform ance scores while maintaining the readability of the models at the same time. We envision that such knowledge could be applied to the development of Human-Centered applications to monitor teams' TMS and provide real-time feedback to improve their performance and affective outcomes on collaborative tasks. For example, an intelligent agent could monitor the interactions of a team performing a brainstorming task and intervene if a decrease in *Specialisation*, *Credibility* or *Coordination* between the members is detected. Previous studies showed that the intervention from an agent playing the role of team leader is perceived to potentially improve the TMS of a team [12, 13]. In this case, if for example a lack of *Coordination* is detected, the agent could mediate the interaction and suggest ways to find a common agreement between the members. The features we found to be the most relevant in estimating TMS dimensions can be easily computed in real-time and be used by the agent to decide when and how to intervene.

Similarly to previous work on automatic analysis of team emergent states, we found that features about turn-taking are also good estimators of TMS. For example, a small number of speaking turns per minute may reflect trust between team members (i.e., they do not need to reply to each other) and so can be used to estimate *Credibility*. A small number of interruptions, which in turn are related to longer average speaking length, may reflect fluid interaction and can therefore be used to estimate *Coordination*. Features related to the spatial arrangements are also good estimators of the TMS dimensions, as they might reflect the tendency of team members to seek (*Credibility*) and provide (*Specialisation*) help according to their expertise, as well as the fluidity of the interaction (*Coordination*). In addition, results show that, in general, by combining multiple modalities (i.e., audio, movement and spatial), we obtain better performances compared to the unimodal and bimodal models, keeping the same number of features.

The difficulty in automatically modelling *Coordination* could be related to the low reliability of the self-reported scores, indicating that this task is difficult also for humans. This result could also be linked to the difficulty for the team members to self-assess *Coordination*, which could be easier estimated by external observers. In the future, we will consider collecting additional annotations of TMS given by external observers.

Our work faces the following limitations. First, the WoNoWa dataset contains a relatively small number of observations, compared to the set of features available. This often occurs when dealing with human behaviour analysis. We show, however, that the 3 dimensions of TMS can be effectively detected as a linear combination of multimodal features. Using simple models such as multiple linear regressions also allowed for interpreting the role of each feature. Second, the generalisability of our findings may be limited to tasks similar to the ones realised in the WoNoWa dataset (i.e., knowledge-based tasks, or "process" tasks according to the classification given in [52]). TMS dimensions could be better modelled using different nonverbal cues in different tasks, such as decision-making or problem-solving. Additionally, the self-reported scores provided by participants show relatively low variability, which is not desirable when running regression models. Finally, we analysed nonverbal behaviour by averaging feature values over large time windows. As a future perspective, our work could be improved by modelling temporal dynamics, for example by computing histograms of co-occurrences [77, 78]. The previous steps of the *Retrieval* phase, which were not considered in this work, could also be included in the analyses to investigate the development of TMS over time.

Acknowledgements The authors would like to thank the reviewers for their useful comments and suggestions.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

References

1. Alborno P, Piana S, Mancini M, Niewiadomski R, Volpe G, Camurri A (2016) Analysis of intrapersonal synchronization in full-body movements displaying different expressive qualities. In: Proceedings of the international working conference on advanced visual interfaces, pp 136–143
2. Argote L, Aven BL, Kush J (2018) The effects of communication networks and turnover on transactive memory and group performance. Organ Sci 29(2):191–206
3. Austin JR (2003) Transactive memory in organizational groups: the effects of content, consensus, specialization, and accuracy on group performance. J Appl Psychol 88(5):866
4. Bachrach DG, Mullins R (2019) A dual-process contingency model of leadership, transactive memory systems and team performance. J Bus Res 96:297–308
5. Beattie GW (1981) Interruption in conversational interaction, and its relation to the sex and status of the interactants. Linguistics 19(1–2):15–36
6. Beyan C, Carissimi N, Capozzi F, Vascon S, Bustreo M, Pierro A, Becchio C, Murino V (2016) Detecting emergent leader in a meeting environment using nonverbal visual features only. In: Proceedings of the 18th ACM international conference on multimodal interaction, pp 317–324
7. Beyan C, Capozzi F, Becchio C, Murino V (2017) Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. IEEE Trans Multimed 20(2):441–456
8. Beyan C, Katsageorgiou V-M, Murino V (2017) Moving as a leader: detecting emergent leadership in small groups using body pose. In: Proceedings of the 25th ACM international conference on multimedia, pp 1425–1433
9. Beyan C, Katsageorgiou V-M, Murino V (2019) A sequential data analysis approach to detect emergent leaders in small groups. IEEE Trans Multimed 21(8):2107–2116
10. Beyan C, Karumuri S, Volpe G, Camurri A, Niewiadomski R (2021) Modeling multiple temporal scales of full-body movements for emotion classification. IEEE Trans Affect Comput
11. Biancardi B, Maisonnave-Couterou L, Renault P, Ravenet B, Mancini M, Varni G (2020) The wonowa dataset: investigating the transactive memory system in small group interactions. In: Proceedings of the 2020 international conference on multimodal interaction, pp 528–537
12. Biancardi B, Giaccaglia I, Ravenet B, Varni G (2021) Virtual leaders supporting the development of transactive memory systems. In: 32e conférence francophone sur l'Interaction Humain-Machine (IHM'20.21). ACM, pp 4–1
13. Biancardi B, O'Toole P, Giaccaglia I, Ravenet B, Pitt I, Mancini M, Varni G (2021) How ECA vs human leaders affect the perception of transactive memory system (TMS) in a team. In: 2021 9th International conference on affective computing and intelligent interaction (ACII). IEEE, pp 1–8
14. Boyd Sr FD (2000) Non-verbal behaviors of effective teachers of at-risk African-American male middle school students. PhD thesis, Virginia Tech
15. Bradski G (2000) The OpenCV library. Dr. Dobb's J Softw Tools
16. Bresin R, Mancini M, Elblaus L, Frid E (2020) Sonification of the self vs. sonification of the other: differences in the sonification of performed vs. observed simple hand movements. Int J Hum Comput Stud 144:102500
17. Campbell N, Sadanobu T, Imura M, Iwahashi N, Suzuki N, Douxchamps D (2006) Multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow. In: LREC, pp 391–394
18. Camurri A, Mazzarino B, Volpe G (2003) Analysis of expressive gesture: the eyesweb expressive gesture processing library. In: International gesture workshop. Springer, Berlin, pp 460–467
19. Canny J (2001) Human-centered computing (technical report). Berkeley, CA: University of California, Berkeley
20. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6(4):284
21. Clarkson E, Day JA, Foley JD (2006) An educational digital library for human-centered computing. In: CHI'06 Extended abstracts on human factors in computing systems, pp 646–651
22. Converse S, Cannon-Bowers JA, Salas E (1993) Shared mental models in expert team decision making. Individual and group decision making: Current Issues 221:221–46
23. Den Y (2018) F-formation and social context: how spatial orientation of participants' bodies is organized in the vast field.

In: Proceedings of LREC 2018 workshop: language and body in real life (LB-IRL2018) and multimodal corpora (MMC2018) joint workshop, pp 35–39

24. Feese S, Arnrich B, Troster G, Meyer B, Jonas K (2011) Detecting posture mirroring in social interactions with wearable sensors. In: 2011 15th Annual international symposium on wearable computers. IEEE, pp 119–120
25. Feese S, Muaremi A, Arnrich B, Troster G, Meyer B, Jonas K (2011) Discriminating individually considerate and authoritarian leaders by speech activity cues. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE, pp 1460–1465
26. Feese S, Arnrich B, Tröster G, Meyer B, Jonas K (2012) Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In: 2012 International conference on privacy, security, risk and trust and 2012 international conference on social computing. IEEE, pp 520–525
27. Forbes RJ, Jackson PR (1980) Non-verbal behaviour and the outcome of selection interviews. J Occup Psychol 53(1):65–72
28. Frid E, Bresin R, Alborno P, Elblaus L (2016) Interactive sonification of spontaneous movement of children-cross-modal mapping and the perception of body movement qualities through sound. Front Neurosci 10:521
29. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit 47(6):2280–2292
30. Grossman R, Friedman SB, Kalra S (2017) Teamwork processes and emergent states. In: The Wiley Blackwell handbook of the psychology of team working and collaborative processes, pp 243–269
31. Hall ET (1966) The hidden dimension, vol 609. Doubleday, Garden City
32. Heldner M, Edlund J (2010) Pauses, gaps and overlaps in conversations. J Phon 38(4):555–568
33. Hollingshead AB (1988) Distributed knowledge and transactive processes in decision-making groups
34. Hollingshead AB (1998) Groupand individual training: the impact of practice on performance. Small Group Res 29(2):254–280
35. Hollingshead AB (2000) Perceptions of expertise and transactive memory in work relationships. Group Process Intergroup Relations 3(3):257–267
36. Hollingshead AB, Brandon DP (2003) Potential benefits of communication in transactive memory systems. Hum Commun Res 29(4):607–615
37. Hung H, Gatica-Perez D (2010) Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Trans Multimed 12(6):563–575
38. Janin A, Baron D, Edwards J, Ellis D, Gelbart D, Morgan N, Peskin B, Pfau T, Shriberg E, Stolcke A et al (2003) The ICSI meeting corpus. In: 2003 IEEE international conference on acoustics, speech, and signal processing, 2003. Proceedings (ICASSP'03), vol 1. IEEE, pp I–I
39. Jayagopi DB, Gatica-Perez D (2010) Mining group nonverbal conversational patterns using probabilistic topic models. IEEE Trans Multimed 12(8):790–802
40. Jayagopi DB, Ba S, Odobez J-M, Gatica-Perez D (2008) Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In: Proceedings of the 10th international conference on Multimodal interfaces. ACM, pp 45–52
41. Jovanovic N, op den Akker R, Nijholt A (2006) A corpus for studying addressing behaviour in multi-party dialogues. Lang Resour Eval 40:5–23
42. Kantharaju RB, Langlet C, Barange M, Clavel C, Pelachaud C (2020) Multimodal analysis of cohesion in multi-party interactions. In: LREC, 2020
43. Kendon A (1990) Conducting interaction: patterns of behavior in focused encounters, vol 7. CUP Archive
44. Kleanthous S, Michael M, Samaras G, Christodoulou E (2016) Transactive memory in task-driven 3d virtual world teams. In: Proceedings of the 9th Nordic conference on human–computer interaction, pp 1–6
45. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163
46. Kozlowski SWJ, Ilgen DR (2006) Enhancing the effectiveness of work groups and teams. Psycho Sci Public Interest 7(3):77–124
47. Kubasova U, Murray G, Braley M (2019) Analyzing verbal and nonverbal features for predicting group performance. In: Proc. Interspeech 2019. ISCA, pp 1896–1900
48. Kwon K, Cho D (2016) How transactive memory systems relate to organizational innovation: the mediating role of developmental leadership. J Knowl Manag
49. Leo FM, González-Ponce I, García-Calvo T, Sánchez-Oliva D et al (2019) The relationship among cohesion, transactive memory systems, and collective efficacy in professional soccer teams: a multilevel structural equation analysis. Group Dyn Theory Res Pract 23(1):44
50. Lepri B, Mana N, Cappelletti A, Pianesi F (2009) Automatic prediction of individual performance from "thin slices" of social behavior. In: Proceedings of the 17th ACM international conference on multimedia, pp 733–736
51. Lewis K (2003) Measuring transactive memory systems in the field: scale development and validation. J Appl Psychol 88(4):587–604
52. Lewis K, Herndon B (2011) Transactive memory systems: current issues and future research directions. Organ Sci 22(5):1254–1265
53. Liang DW, Moreland R, Argote L (1995) Group versus individual training and group performance: the mediating role of transactive memory. Personal Soc Psychol Bull 21(4):384–393
54. Liao J, Jimmieson NL, O'Brien AT, Restubog SLD (2012) Developing transactive memory systems: theoretical contributions from a social identity perspective. Group Organ Manag 37(2):204–240
55. Marks MA, Mathieu JE, Zaccaro SJ (2001) A temporally based framework and taxonomy of team processes. Acad Manag Rev 26(3):356–376
56. Martínez-Cagigal V (2018) Sample entropy. https://it.mathworks.com/matlabcentral/fileexchange/69381-sample-entropy. Accessed 18 Mar 2021
57. Mawalim CO, Okada S, Nakano YI, Unoki M (2023) Personality trait estimation in group discussions using multimodal analysis and speaker embedding. J Multimodal User Interfaces 1–17
58. Michinov E (2007) Validation de l'échelle de mémoire transactive en langue française et adaptation au contexte académique. Revue Européenne de Psychologie Appliquée/Eur Rev Appl Psychol 57(1):59–68
59. Michinov E, Olivier-Chiron E, Rusch E, Chiron B (2008) Influence of transactive memory on perceived performance, job satisfaction and identification in anaesthesia teams. Br J Anaesth 100(3):327–332
60. Miller PW (1988) Nonverbal communication. what research says to the teacher. ERIC
61. Moreland RL (1999) Transactive memory: learning who knows what in work groups and organizations. In: Thompson L, Messick D, Levine J (eds) Shared cognition in organizations: the management of knowledge
62. Moreland RL (2010) Are dyads really groups? Small Group Res 41(2):251–267
63. Moreland RL, Myaskovsky L (2000) Exploring the performance benefits of group training: transactive memory or improved communication? Organ Behav Hum Decis Process 82(1):117–133

64. Moreland RL, Thompson L (2006) Transactive memory: learning who knows what in work groups and organizations. In: Small groups: key readings. Psychology Press, New York, pp 327–346
65. Murata K (1994) Intrusive or co-operative? A cross-cultural study of interruption. J Pragmat 21(4):385–400
66. Nanninga MC, Zhang Y, Lehmann-Willenbrock N, Szlávik Z, Hung H (2017) Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 206–215
67. Oertel C, Funes MKA, Sheikhi S, Odobez J-M, Gustafson J (2014) Who will get the grant? A multimodal corpus for the analysis of conversational behaviours in group interviews. In: Proceedings of the 2014 workshop on understanding and modeling multiparty, multimodal interactions, pp 27–32
68. Pavitt C (2003) Colloquy: do interacting groups perform better than aggregates of individuals? Why we have to be reductionists about group memory. Hum Commun Res 29(4):592–599
69. Peltokorpi V, Hood AC (2019) Communication in theory and research on transactive memory systems: a literature review. Top Cognit Sci 11(4):644–667
70. Rahimpour M (2014) The nature of transactive memory systems in emergency medicine teams based on observations and communication analysis. PhD thesis, Carleton University
71. Rapp T, Maynard T, Domingo M, Klock E (2021) Team emergent states: what has emerged in the literature over 20 years. Small Group Res 52(1):68–102
72. Ren Y, Argote L (2011) Transactive memory systems 1985–2010: an integrative framework of key dimensions, antecedents, and consequences. Acad Manag Ann 5(1):189–229
73. Sanchez-Cortes D, Aran O, Mast MS, Gatica-Perez D (2010) Identifying emergent leadership in small groups using nonverbal communicative cues. In: International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction, pp 1–4
74. Sanchez-Cortes D, Aran O, Mast MS, Gatica-Perez D (2011) A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Trans Multimed 14(3):816–832
75. Sanchez-Cortes D, Aran O, Gatica-Perez D (2011) An audio visual corpus for emergent leader analysis. In: Workshop on multimodal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI. Citeseer
76. Sanchez-Cortes D, Aran O, Jayagopi DB, Mast MS, Gatica-Perez D (2013) Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. J Multimodal User Interfaces 7(1–2):39–53
77. Van Segbroeck M, Van hamme H (2009) Unsupervised learning of time-frequency patches as a noise-robust representation of speech. Speech Commun 51(11):1124–1138
78. Van hamme H (2008) Hac-models: a novel approach to continuous speech recognition. In: Ninth annual conference of the international speech communication association
79. Varni G, Mancini M (2020) Movement expressivity analysis: from theory to computation. In: Modelling human motion. Springer, Berlin, pp 213–233
80. Walocha F, Maman L, Chetouani M, Varni G (2020) Modeling dynamics of task and social cohesion from the group perspective using nonverbal motion capture-based features. In: Companion publication of the 2020 international conference on multimodal interaction, pp 182–190
81. Wegner DM (1987) Transactive memory: a contemporary analysis of the group mind. In: Theories of group behavior. Springer, Berlin, pp 185–208
82. Yoo Y, Kanawattanachai P (2001) Developments of transactive memory systems and collective mind in virtual teams. Int J Organ Anal 9(2):187–208
83. Zhong X, Huang Q, Davison RM, Yang X, Chen H (2012) Empowering teams through social network ties. Int J Inf Manag 32(3):209–220
84. Zhou Z, Pazos P (2020) Empirical perspectives of transactive memory systems: a meta-analysis. Team Perform Manag Int J