frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fcomp.2023.1062342

OPEN ACCESS

EDITED BY
Chloé Clavel,
TELECOM ParisTech, France

REVIEWED BY
Massimo Moneglia,
University of Florence, Italy
Kazutaka Shimada,
Kyushu Institute of Technology, Japan

*CORRESPONDENCE
Philippe Blache
✉ blache@ilcb.fr

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 05 October 2022
ACCEPTED 06 February 2023
PUBLISHED 02 March 2023

CITATION
Pellet-Rostaing A, Bertrand R, Boudin A, Rauzy S and Blache P (2023) A multimodal approach for modeling engagement in conversation. *Front. Comput. Sci.* 5:1062342. doi: 10.3389/fcomp.2023.1062342

A multimodal approach for modeling engagement in conversation

Arthur Pellet-Rostaing[1,2], Roxane Bertrand[1,2], Auriane Boudin[1,2], Stéphane Rauzy[1,2] and Philippe Blache[1,2]*

[1]Laboratoire Parole and Langage (LPL-CNRS), Aix-en-Provence, France, [2]Institute of Language, Communication and the Brain (ILCB), Marseille, France

Recently, engagement has emerged as a key variable explaining the success of conversation. In the perspective of human-machine interaction, an automatic assessment of engagement becomes crucial to better understand the dynamics of an interaction and to design socially-aware robots. This paper presents a predictive model of the level of engagement in conversations. It shows in particular the interest of using a rich multimodal set of features, outperforming the existing models in this domain. In terms of methodology, study is based on two audio-visual corpora of naturalistic face-to-face interactions. These resources have been enriched with various annotations of verbal and nonverbal behaviors, such as smiles, head nods, and feedbacks. In addition, we manually annotated gestures intensity. Based on a review of previous works in psychology and human-machine interaction, we propose a new definition of the notion of engagement, adequate for the description of this phenomenon both in natural and mediated environments. This definition have been implemented in our annotation scheme. In our work, engagement is studied at the turn level, known to be crucial for the organization of the conversation. Even though there is still a lack of consensus around their precise definition, we have developed a turn detection tool. A multimodal characterization of engagement is performed using a multi-level classification of turns. We claim a set of multimodal cues, involving prosodic, mimo-gestural and morpho-syntactic information, is relevant to characterize the level of engagement of speakers in conversation. Our results significantly outperform the baseline and reach state-of-the-art level (0.76 weighted F-score). The most contributing modalities are identified by testing the performance of a two-layer perceptron when trained on unimodal feature sets and on combinations of two to four modalities. These results support our claim about multimodality: combining features related to the speech fundamental frequency and energy with mimo-gestural features leads to the best performance.

KEYWORDS

engagement model, multimodality, conversational skills, conversational agents, engagement classification, annotated corpora

1. Introduction

The notion of engagement has gained increasing attention over the recent years, especially from research focusing on human-agent interactions (Oertel et al., 2020). This trend is fueled by the need to design artificial agents capable of adapting their actions according to users' behavior. Many works have stressed the necessity to evaluate artificial social agent's performances beyond their ability to produce linguistically appropriate

responses, but also by taking users' experience into account (Dybala et al., 2009; Bickmore et al., 2010). The notion of engagement is then of great importance, in particular because of its direct impact to the length and the success of an interaction (Sidner et al., 2004, 2005; Khatri et al., 2018; Venkatesh, 2018; Oertel et al., 2020). In spite of their important role, neither conversation length nor agents' success at influencing users constitute the only metrics for the success of an interaction. Rendering possible and efficient a conversation lies in building a shared knowledge (called *common ground*) by participants (Clark, 1996): the viability of any verbal interaction is built upon a certain level of collaboration between the participants which is implemented through various acts of grounding. A whole array of behaviors necessary for the interaction to go smoothly (e.g., feedbacks, gestures) requires a certain level of cognitive investment, in addition to those necessary for production and comprehension. Evaluating engagement is then a way to explain agents' ability to achieve such goals (Novielli et al., 2010; Anzalone et al., 2015). In this perspective, engagement can be viewed as a variable regulating the level of collaboration and more generally the interaction dynamics, beyond its initiation and its duration.

While a substantial research effort has been done to automatically estimate engagement in the context of human-agent interactions (Oertel et al., 2020) only few studies have looked at variation in the degree of engagement during casual human-human interactions. Technically, most of the latter are based on visual features, such as facial *Action Units* activation, intensity and gaze direction (Dermouche and Pelachaud, 2019), facial landmarks and texture-based features (Huang et al., 2016), gestures (Dermouche and Pelachaud, 2018), etc. However, the use of *Action Units* (AUs) and complex face-related features can make the interpretation of the results difficult because involving many different parameters, rendering them often difficult to categorize. It is then necessary to identify a subset of AU relevant in the case of engagement. This calls for attempting to classify engagement level on the basis of more readily interpretable features. Moreover, multimodality has been little explored in human-human conversation except in Fedotov et al. (2018), proposing a binary engagement-disengagement classification. Although valuable, such models remain limited in the sense that variation in engagement level is more subtle than what is implied by a binary scale.

We propose to model engagement in human-human conversations by using (1) a multimodal set of interpretable features and (2) a detailed multiclass scale. To this end, we have created a new resource, based on a 4.5 hours conversational corpus enriched with the level of engagement on a 5-level scale. Morpho-syntactic, prosodic, mimo-gestural features and engagement have been annotated at the turn level. To the best of our knowledge, the only approach performing an automatic estimation of engagement at the turn level (applied to a corpus of conversations between students and a virtual tutoring agent) has been presented in Forbes-Riley et al. (2012). In our work, we propose to apply a comparable approach to human-human conversations. This choice is motivated by the fact that turns represent one of the relevant units for accounting the structural organization of conversations (Sacks et al., 1974). We also propose to use a more detailed scale than the one used by Forbes-Riley et al. (2012) who restricted their classification task to a binary problem (engagement vs. disengagement). Finally, another originality is that our work aims at classifying engagement level when participants are holding the turn, thus avoiding to segment more or less arbitrarily the input signal as it is the case with many other works that do not distinguish between the participant's role (speaker or listener). This is possible thanks to the manual annotation of feedbacks performed previously (Boudin et al., 2021).

We present in this paper a new annotated resource for studying engagement in interaction, together with tools and methods offering a way to address this question in the more general case: *unrestricted conversations*. Engagement being not really well defined in the literature, we propose in the second section a large survey of this question both in human-machine and human-human environments. Section 3 presents dataset and features selection. Results are described and discussed in Sections 4 and 5.

2. Defining engagement

In spite of the fact that engagement is broadly used in the literature, this notion still relies on imprecise and often informal definitions. We propose in this section to give an overview of this question, highlighting its main features in the perspective of elaborating a definition based on explicit and quantifiable characteristics.

2.1. Engagement in the literature

One of the most influential definitions of engagement has been presented in Sidner and Dzikovska (2002) who characterize engagement as "*the process by which two (or more) participants establish, maintain, and end their perceived connection to each other during an interaction*." This definition, elaborated in a human-machine interaction context, aims at identifying the most important factors by which an embodied conversational agent (ECA) can make the conversation efficient. Several works are based on this definition either to annotate engagement (Yu et al., 2004; Hsiao et al., 2012; Leite et al., 2015) or to automatically detect it (Ooko et al., 2011; Huang et al., 2016; Foster et al., 2017; Ben-Youssef et al., 2019). However, this definition raises three limitations. First, engagement is considered as a process and rendering difficult to propose a measure for estimating its level varying along the interaction. It is also limited by a too global scope. Three major goals are presented: *initiating*, *maintaining* or *ending* an interaction, excluding subtle variations of engagement such as showing interest or affiliation. This approach resumes then engagement as a binary process with two states: *engagement* or *disengagement*. Second, this definition excludes the idea that engagement involves both interlocutors. The level of each interlocutor is considered independently from the other. But conversations are contextual phenomena, where the interlocutors have a mutual influence and adapt their behaviors according to their partner. From this point of view, looking for a global level of engagement including both participant adjustment seems difficult. Third, identifying engagement with the unfolding of the interaction itself and not as a variable influencing this unfolding makes unclear why the notion of engagement should be mentioned at all. Overall,

the different works relying on the definition proposed in Sidner and Dzikovska (2002) end up in considering engagement as a *state*, taken globally (Ooko et al., 2011). This seems to be too general, without leaving place to a more fine-grained description, based on the identification of different features characterizing this notion.

On its side, Poggi (2007) defines engagement as *the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing the interaction*. This state-based definition has two main interests. First, engagement is considered as an individual property and at the same time offers the possibility to include both participants by evaluating possible discrepancies between their levels of engagement. Second, it allows to consider that engagement fluctuates over time and can hold different levels. Moreover, in contrast with Sidner and Dzikovska (2002), the work presented in Poggi (2007) takes into consideration the mutual intention of participants to collaborate, making successful the interactional goals achievement. Engagement is here an attitude in which both the context and the content of the conversation interact. Nonetheless, no specific interactional behaviors are pointed out, making unclear the relative importance of particular actions or cues by which engagement is implemented.

Peters et al. (2005a) proposes to define engagement as *"an emotional state linked to the participant' goal of receiving and elaborating new and potentially useful knowledge"*. This definition includes an emotional dimension, which is absent from Sidner and Dzikovska (2002) and only implicit in Poggi (2007). The *common ground* (denoted as knowledge) is put in relation with engagement and information flow, interacting as a common goal of the participants. This makes Peters et al. (2005a) definition truly specific to the conversational context. Moreover, it allows to understand that being engaged is not only about attributing a high value to interlocutor's engagement or to what is being discussed, but also about the wish to contribute to the conversation. In other words, as underlined previously, engagement is correlated with the willingness to invest effort in the interaction. A limitation to this conceptualization however is that speakers can be strongly engaged while not necessarily elaborating new knowledge, as it is for example the case when two people have a heated argument and keep repeating the same information. Since participants' goals can go beyond informativeness it seems more appropriate to conceptualize engagement as a function of participants' goals in general.

Finally, in a human-machine interaction perspective, the notion of engagement is used to improving agents' conversational abilities. Several works have focused on designing systems able to recognize users' intention to initiate an interaction (Bohus and Horvitz, 2009; Foster et al., 2017) or to detect when the user is likely to step away from the interaction (Nakano and Ishii, 2010). Generally speaking, even though the notion of engagement is widely used in the context of human-machine interaction (typically for embodied conversational agents), an explicit definition is still lacking or imprecise in most of these works (Mower et al., 2007; Bednarik et al., 2012; Bonin et al., 2012; Leite et al., 2013; Baker et al., 2014). In some works, engagement refers to the influence of an agent on the user (Novielli, 2009), for some others to the spatial distance between the virtual agent and the user (Michalowski et al., 2006).

Overall, these definitions highlight two main components of engagement: attention and emotion. Attention to the interlocutor or to interaction-relevant objects is often put forward as a constitutive characteristics (Sidner et al., 2004; Yu et al., 2004). Inattention can indeed be considered as a sign of reduced investment in contributing to the dynamics of the conversation. On the opposite, sustained attention to different stimuli at appropriate times is also a prerequisite for an affective involvement (Peters et al., 2009). The relationship between emotion and attention is bidirectional since emotionally relevant stimuli capture attention (Peters et al., 2009). Therefore attention can be viewed as a minimal form of engagement. The role played by emotions in engagement can be inferred from Poggi (2007) definition as the value attributed to interacting with one's interlocutor is likely of emotional nature.

As a result, none of these definitions cannot be taken beyond its specific context and completely fulfill the goal of estimating an engagement level during an interaction. Cumulative elaboration on the concept of engagement makes necessary the elaboration of a global definition, gathering the different characteristics of the existing ones and offering solution for the elaboration of a computational model.

2.2. Concepts and notions in relation with engagement

One way to thoroughly flesh out the concept of engagement is to look at how it differs from related notions. This endeavor is also necessary to justify the very use of this concept.

A notion sometimes equated with engagement is *interest* (Yu et al., 2004). This reduction makes sense under the light of (Poggi, 2007) definition of engagement in terms of value. Other authors highlight the role played by preferences in shaping the emotional relevance of conversation topics (Glas and Pelachaud, 2015a). The main issue with identifying engagement to interest is that the latter does not specify any attitude toward the unfolding of the interaction. It is even possible to imagine situations in which interest co-occurs with disengagement.

Another notion close to that of engagement is *rapport*, defined as a mix of "*mutual attentiveness, positivity and coordination*" (Tickle-Degnen and Rosenthal, 1990). Both notions point to the collaborative aspect of conversation. *Rapport* however is something that is built up in the long term, while engagement rather relates to the short term. Moreover, a crucial difference relates to the emotional dimension. While rapport is partly defined by affects valence, the affective aspect of engagement solely relates to the arousal dimension. Experiencing rapport implies a positive attitude toward the interlocutor, whereas being engaged can result from being mad or distressed.

Finally, the concept of *involvement* is often used interchangeably with engagement. Some works explicitly advocate for abandoning any distinction between the two (Skarbez et al., 2017). In this way, Bickmore et al. (2010) conceptualize engagement as "*The degree of involvement a user chooses to have with a system over time*." This is not surprising given the notorious characterization of involvement as the "*consequence of focusing one's energy and attention on a coherent set of stimuli*" (Witmer and Singer, 1998). Both notions relate to the investment of efforts into

a specific activity. When applied to social interactions the concept of involvement is arguably very close to that of engagement.

2.3. Ground rules for a proper definition of engagement

An adequate definition of engagement in the perspective of a modeling must satisfy different requirements. First, it must be valid to both human-agent and human-human interactions. This prohibits using the term "user" in the definition itself, even though one can mention it when the context allows it (Glas and Pelachaud, 2015b). This also incites not to consider engagement only on the basis of having the interaction with an ECA going on for as long as possible (Oertel et al., 2020). Second, it must allow studying heterogeneity and co-influence of participants' degree of engagement. Therefore, engagement must be defined at the individual level (which does not mean not taking into account both participants, as explained above). Third, the definition must take into account the importance of participants' goals in the interactional context. Indeed, the notion of engagement is motivated by the need to explain why a conversation can be successful or not. As already pointed out, the success of a conversation is partly defined by participants' respective goals as well as by the value they place on these goals. Of course, interlocutor's degree of engagement might favor the ability to achieve one's goals, but this should not be considered as something that defines engagement. Fourth, the definition must be specific enough to avoid encompassing other types of engagement. Conversational engagement has to be distinguished from task engagement for example (Oertel et al., 2020). Last, but not least, any proper definition of engagement must highlight its two core components, namely attention and emotion.

2.4. Features and methods for characterizing engagement

This section presents the different features used for engagement classification and prediction. Both tasks are described in human-agent and human-human interactions, even if greater attention is paid to the latter. Moreover, we focus on intra-conversation variation in the degree of engagement. We do not consider here studies related to the recognition of intentions to engage in conversation, to the analysis of artificial agents' conversational performance or to conversational system design.

Facial expressions are by far the most exploited features in the literature by taking facial *Action Units* intensity and activation as input (Dhamija and Boult, 2017; Liu and Kappas, 2018; Ben-Youssef et al., 2019; Dermouche and Pelachaud, 2019). Smiles are also informative about the level of engagement (Allwood and Cerrato, 2003; Castellano et al., 2009; Leite et al., 2015). Supplying raw facial images to a convolutional neural network has also been shown to yield excellent results (0.92 F-score) (Huang et al., 2016).

Gaze is another common feature used for automatic engagement estimation. A variety of gaze-related features have been used, such as the time spent at looking at a robot (Castellano et al., 2009), gaze transition 3-gram sequences (Ishii et al., 2013) or number, duration and amplitude of saccades (Bednarik et al., 2012). Looking at the interlocutor is hypothesized to be a cue of engagement, and looking around to be a cue of disengagement (Sidner et al., 2004). However, looking at conversation-relevant objects is considered as indicative of engagement, especially when the current interaction involves a collaborative task (Sidner et al., 2004). From the listener's perspective gaze is also thought to be a way to provide feedbacks, signaling interlocutor's level of engagement (Peters et al., 2005b).

Engagement classification has also been performed on the basis of *prosodic* features. In general, this information is represented by vectors of low-level features (f0, MFCCs, log-energy, etc.) extracted from the speech signal within sliding windows. Functions are applied to these to produce mid-level vectors summarizing speech prosodic characteristics (Yu et al., 2004; Hsiao et al., 2012). Speech prosody is considered as cue speakers' level of arousal (Yu et al., 2004). The use of such features is consistent with emotional investment being at the core of the concept of engagement.

Other verbal features (e.g., syntactic complexity) are rarely involved in the models, especially in tandem with nonverbal features. Such features were used by (Forbes-Riley et al., 2012). However, they are too study-specific (e.g., student's response incorrectness, use of a remediation question by the artificial agent) to be used in other works.

Exploiting *multimodality* is arguably advantageous when working on conversational engagement. Indeed, both facial expressions, gestures and prosody can convey information about the participant's emotional state. Gaze and gestures may also cue investment in grounding information. Combining features from different modalities improves classification performance. Fedotov et al. (2018) reports improvement when combining auditory and visual (e.g., lips movements, gaze) features. Likewise, with a LTSM model, Dermouche and Pelachaud (2019) obtain better results when taking gaze, head position, AUs activation and intensity as input compared to when being trained on each feature separately. The performance of users' engagement breakdown detection is also improved when using multimodal features (Ben-Youssef et al., 2019).

Few studies have assessed the relative importance of different modalities when classifying or predicting the level of engagement. Huang et al. (2016) have found that training SVMs on facial texture features, facial landmarks and head pose yielded better results than training them with acoustic features. Even better results were obtained by simply feeding raw facial images to a convolutional neural network. The importance of facial expressions have been established by other studies. The LTSM model for engagement prediction proposed in Dermouche and Pelachaud (2019) performs better when trained on Action Units intensity and activation than when trained on gaze direction and head pose respectively. Combining both sets of features leads to only marginally better results compared to AUs alone. Fedotov et al. (2018) evaluate modalities contributions by looking at their model's performance for all combinations of two to four features. Lips-based and acoustic features were found to be the most informative modalities. Contrasting with the aforementioned studies, facial expressions contributed only slightly to classification. The limited number of works investigating modalities relative contributions

as well as the presence of inconsistent findings calls for further exploration of this issue. This endeavor represents a critical step toward a proper characterization of engagement.

In general, most human-agent studies focus on discriminating engagement from disengagement, with the further goal of improving systems' ability to quickly detect and counteracting engagement breakdown (Liu and Kappas, 2018; Ben-Youssef et al., 2019). On the contrary, human-human studies tend to work with a more detailed engagement scale made of five levels. Despite the greater number of classes they manage to achieve very high classification scores, reaching 0.99 F-score (Dermouche and Pelachaud, 2019).

2.5. Definition of engagement used in this study

Based on the above review of engagement definitions, we propose to define this concept as a state of attentional and emotional investment in contributing to the conversation by processing partner's multimodal behaviors and grounding new information.

Engagement takes into account the value participants attribute to their goals, and is subject to variations over the course of the interaction (as participants' goals and expectations may shift according to topic changes, the introduction of new information, etc.).

This paper focuses on the first part of this definition by proposing a predictive model of engagement at the turn level. This information is necessary before addressing the two next questions, namely engagement variations and adequacy with the speaker's goals.

3. Modeling engagement

We present in this section an experiment based on an original dataset for modeling engagement. We first describe a specific methodology for annotating manually the ground truth by proposing an original segmentation (based on turns) and the tools for extracting the multimodal features to be involved in the model. We then apply different machine learning methods for building a predictive model of the engagement level.

3.1. Dataset

The dataset is built from the corpus Paco-Cheese (Amoyal et al., 2020; Priego-Valverde et al., 2020), containing audio-video recordings of natural dyadic face-to-face interactions between native French speakers. The corpus is made of 31 conversations, lasting between 15 and 20 mns. Participants were instructed to read a short humorous story before engaging in an unconstrained conversation. The corpus is fully transcribed and enriched with various annotations. Interpausal units (IPUs), tokens, laughter and pauses were semi-automatically transcribed and aligned onto the audio signal thanks to the SPPAS software (Bigi, 2012). Our annotation procedure involves segmenting the audio-visual recordings according to the conversation status (speaker/interlocutor) of each participant. As explained below, the segmentation is based on IPUs and feedbacks time boundaries. The annotation of feedbacks time location and type has been previously performed on 14 dyads (Boudin et al., 2021). This subset of the Paco-Cheese corpus will constitute our dataset, which includes approximately 8 h of recordings.

3.2. Engagement annotation

The originality of our work is twofold: (1) segmenting the input into turns instead of arbitrary segments and (2) annotating engagement level into a 5-levels scale instead of a binary one. We present in this section our proposals addressing these aspects.

3.2.1. Turns detection

In the literature, the degree of engagement is usually indicated on annotator-defined segments: annotators determine segments boundaries within which engagement level is considered homogeneous. A new segment is defined when this level changes. We adopt a more generic and regular segmentation by annotating the degree of engagement at the turn-level. This choice is first motivated by the fact that turns correspond to a basic unit with a certain semantic, lexical and syntactic coherence. They also form the backbone of the interactional structure. We hypothesized that they can define time intervals within which engagement level can be uniform. Moreover, this choice also facilitates the use of a multimodal set of features, by assuming that interaction and synchronization between the cues from different modalities can be done at this level.

Restricting classification to moments when the participant is holding the turn renders possible to combine mimo-gestural and prosodic features, which is more problematic when using annotator-defined segments (such segments might contain both moments when the participant is speaking or listening, making prosodic data not always available). This may be one of the reasons why the feature set remains unimodal in many studies. Distinguishing intervals on which the participant is either speaker or listener is also a way to accurately pinpoint features importance. For example, Fedotov et al. (2018) found that lips movement are of high relevance, but authors also note that this result may be due to the fact that annotators perceive participants as more engaged when they are speaking. This problem of interpretation is alleviated by taking separately intervals when the participant is speaking or listening. Finally, another important aspect is that features relevance for engagement classification may differ as a function of the participant's conversational role. Some studies have taken this into account by adding a binary variable indicating whether the participant is speaking or not in their model (Dermouche and Pelachaud, 2018). But this does not allow contrasting models of speaker's and listener's engagement. Our study is limited to the automatic estimation of engagement level when participants are holding the turn.

Technically, turns boundaries were automatically detected using a Python module (PyElan) that connects the ELAN annotation software with Python script especially designed for this study. This module allows parsing, modifying and creating

ELAN files. It includes other functionalities such as extraction of speech activity or feedbacks time boundaries. Turn detection is based on the IPUs (defined as block of speech bounded by silent pauses of 200 ms), excluding IPUs composed by feedbacks only. The detection procedure is as follows: whenever the participant produces an IPU, the turn is attributed to her unless the IPU corresponds to a feedback. This restriction is motivated by the fact that feedback production is by definition an activity of the listener. In simple (but not so trivial) words, the participant holds the turn for as long as her interlocutor does not take it. Consistent with the approach adopted by several studies among them (Gravano and Hirschberg, 2011), IPUs are collapsed all together until the next speaker change. The turns identified by the algorithm correspond to sequences of alternating IPUs and pauses, the latter being considered as akin to *Transition Relevance Places* (Sacks et al., 1974). A turn ends when the main speaker is pausing while her interlocutor takes the turn (as defined above). However, the turn does not end if the main speaker is still speaking: in this case of overlap, the turn is allocated to both participants.

Turns shorter than one second were discarded because of the difficulty of annotating them as well as the risk of extracting unreliable features on such short intervals. The number of turns shorter than 1 s in the whole dataset amounts to 1,040. Overall, a total of 1,200 turns has been kept.

3.2.2. Annotation procedure

The degree of engagement was manually annotated by two experts. The dataset being already existing, it has not be possible to ask the speakers for a self-reporting. However, this type of annotation is interesting if we think important to take into account the goals and values of the speaker, which is less relevant in the type of interaction (narration) we are working on.

At this stage, we decided to annotate engagement for the speaker's production only, not the listener's one, which is limited to feedback. Of course, the number and the type of feedback produced by the listener may have a consequence on the speaker's production. We chose to focus on speakers only for two main reasons. First, a preliminary analysis of the corpus have shown a regular feedback production across the different listeners (Boudin et al., 2021). Second, and more importantly, it is not possible to assess a direct engagement value to a feedback, which are complex objects with different functions. At the annotation level, we therefore decided to only annotate each turn, independently from the rest of the interaction. This choice does not have any impact on the interpretation of the results, in particular the analysis of a potential correlation between the engagement levels of the participants along the interaction.

Following previous studies on engagement in human-human conversations (Yu et al., 2004; Huang et al., 2016; Dhamija and Boult, 2017; Dermouche and Pelachaud, 2019), we define a 5 levels scale for annotating engagement which outperform the standard binary classification. For each turn, annotators had to select one label among:

- Level 1: strongly disengaged.
- Level 2: disengaged.
- Level 3: neutral.
- Level 4: engaged.
- Level 5: strongly engaged.

Annotations were made using ELAN software. For each speaker in the corpus, an ELAN file was created with the segments corresponding to the speaking turns, which means that annotators visualize this information in the corresponding Elan layer. For each segment, the annotators had to select an engagement value between 1 and 5 (as detailed above). Even though annotators were experts, we decided to keep generic instructions and the assignment of the engagement value a holistic manner. A final adjudication has been done between them to decide in case of conflict.

Note that we decided to chose a classical 5-level scale in order to obtain more information in comparison with a binary classification. This granularity seems to us relevant, in particular because it allows to compare different granularities by merging different classes. We present in the experiment the results comparing 3-level to 5-level classifications.

In terms of annotation, here are the guidelines given to the annotators in order to assess the engagement level:

- How willing is the participant to contribute to the progress of the conversation?
- How invested is the participant in what she is saying?
- How interested is she in the conversation?

These questions were designed so as to be the most coherent with our definition of engagement without being too theoretically sophisticated. As it is usually the case with this type of instructions, they stay at a general level, but also partially redundant. In our procedure, we proposed to distinguish between different types of information: the actions of the participants in the quality and the success of the interaction, the intensity of his/her actions (in terms of investment) and the quality of his/her reactions, showing an interest. In the annotators' instructions, we voluntarily do not to point out specific behaviors (e.g., smiling) in order to avoid a possible bias in the modeling by inflating their relevance when automatically estimating engagement. Before performing the task, annotators were presented with short videos of both strongly disengaged and strongly engaged turns in order to make them aware of the whole range of variation of engagement level. They were also asked to leave out any turn for which annotating was too troublesome because of short duration or because of the presence of too many overlaps, rendering the participant's speech hardly audible.

3.3. Feature extraction

Our model relies on a limited set of features from verbal and non-verbal modalities: acoustic, mimogestural and morpho-syntactic cues. Note that some features have been annotated manually (gesture intensity and feedbacks) and that all

annotations have been manually corrected. At this stage, this is of course a limitation when trying to implement the level assessment automatically. However, these restrictions are due in the first case to constraint coming from the quality and the framing of the video and in the second case to the absence of automatic tool for efficiently recognizing a feedback. We are working on these two aspects by acquiring new datasets and developing new tools.

3.3.1. Prosodic features

The values of the fundamental frequency (f0) over the course of the interaction were extracted using Praat (Boersma and Weenink, 1996). The time series of f0 values was segmented according to turns. Values belonging to particular turn were concatenated into a single vector. Functionals (max, min, mean, standard deviation, and inter-quartile range) were then applied, producing a vector of length 5 summarizing f0-related information. On its side, energy was directly extracted from the audio signal. For all non-silent regions of a turn, energy was computed within overlapping 40 ms windows, with a stride value set to 20 ms. Following works exploiting prosodic information (Hsiao et al., 2012; Ben-Youssef et al., 2019), energy values were log-transformed. Information contained in the resulting vectors was finally summarized in the same way as for the f0. Features related to the root mean square deviation (RMSE) were extracted following the same procedure. For some turns it was not possible to obtain a value of the f0 due to their short duration. These turns were removed when classifying engagement level. Based on previous works (Oertel et al., 2011) our hypothesis was that *higher pitch and speech signal energy are positively associated with engagement*.

Features related to aspects of speech duration are also included in the model. For each turn, the turn-taking delay is extracted. It corresponds to the elapsed time before the participant starts speaking once her interlocutor has effectively released the turn, and the fraction of time the speaker was pausing during a given turn. The articulation rate, based on the annotation of syllable boundaries automatically detected with SPPAS, is computed by dividing the number of syllables by the total time spent to produce these syllables (thus excluding silent pauses). We hypothesize that *articulation rate cues a higher level of engagement*. Our second hypothesis is that *turn-taking delay and pauses are negatively associated with engagement*, in accordance with the turn taking system based on the minimization of too long delay between turns and too long pauses or gaps (Sacks et al., 1974; Levinson and Torreira, 2015).

To summarize, and according to the literature, speaking more rapidly, with a higher pitch and energy, and taking the turn rapidly could be cues for a higher engagement level.

3.3.2. Mimo-gestural features

Gestures and expression have been shown to play a central role in the analysis of participants behaviors. In the case of our study, we have decided to focus on two types of such elements: smiles and head nods. We do not, at this stage, take into account action units first because think important to limit the number of features to those directly in connection with the type of phenomenon we want to study and second for a purely technical reason. The positions of the participants in the video do not make it possible to extract automatically in a precise manner the main action units. The same restriction applies to gaze. A manual annotation of AUs should then have been necessary. Taking into account the time and budget constraints of this research, in spite of their potential role in the engagement modeling (in particular gaze directions), we decided then not to add AU in the model and to limit mimo-gestural features to smiles and nods.

Semi-automatic annotation is carried out using SMAD (Amoyal and Priego-Valverde, 2019; Rauzy and Amoyal, 2020; Boudin et al., 2021). Annotations provide information about both smiles duration and intensity, from S0 (no smile) to S4 (laughing smile). For each turn, the relative duration of level of smile intensity was computed, yielding a vector of length 5.

Nods frequency is calculated by dividing the number of head nods by the turn duration. Moreover, arms and hands gestures are included in the model. These annotations has been restricted to manually indicating whether gestures intensity was low, medium or high. Gestures intensity is understood as comprising both the quantity of arms-hands movements and their amplitude. Two expert annotators carried out this task. They were instructed to annotate the intensity level only by taking into account frequency and amplitude into the turn, with respect to the rest of the interaction. As for action units, technical video limitation of our corpus did not allow an automatic estimation of this information. An adjudication has been done in the case of conflict by annotators. Our third hypothesis is that *the greater smiles, nods and gestures intensity, the higher the degree of engagement*.

3.3.3. Morpho-syntactic features

Lemmas and Part-of-Speech were extracted from tokens transcription thanks to the MarsaTag analyser (Rauzy et al., 2014). The fourth hypothesis is that *a greater level of engagement would be associated with an increased discourse structure complexity*. The syntactic richness of each turn is approximated by the frequency of conjunctions and modifiers (i.e., adverbs and adjectives). This approximation is based on the fact that modifiers introduce richer and more complex information by introducing new embedded constituents. This measure has been first proposed in Blache et al. (2020).

Note that other type of morpho-synactic cues have not been considered for this model, in particular turn length. The reason is that the type of the interaction (narrative-like) intrinsically favors longer turns, without being an indication of any engagement.

3.4. Features selection

In order to reduce feature space dimensionality, a subset of features is kept when performing classification. Features were selected based on the strength of their correlation with the degree of engagement, as measured by Pearson correlation coefficient (see Table 1). Note that feature selection has been applied to the entire dataset, instead of the training subset because of using nested cross-validation, which limits the risks of over-fitting (see next section). Only features for which the absolute value of the coefficients was

TABLE 1 The coefficient of correlation ρ (Pearson) between the degree of engagement and the various features proposed in our analysis.

Feature	ρ	Feature	ρ
Turn duration	0.09	f0 (max)	0.16
No smile (S0)	−0.26	f0 (mean)	0.22
Closed mouth smile (S1)	0.08	f0 (inter-quartile range)	0.24
Open mouth smile (S2)	0.01	Log-energy (min)	0.18
Wide open mouth smile (S3)	0.16	Log-energy (max)	0.34
Laughing smile (S4)	0.28	Log-energy (mean)	0.28
Head nods frequency	0.13	Log-energy (std)	0.04
Gestures intensity	0.30	Log-energy (inter-quartile range)	−0.001
Conjunctions frequency	−0.001	RMSE (min)	0.19
Modifiers frequency	−0.09	RMSE (max)	0.39
Turn-taking delay	−0.16	RMSE (mean)	0.39
Pauses	−0.08	RMSE (std)	0.41
f0 (min)	−0.02	RMSE (inter-quartile range)	0.38

higher than 0.1 were retained. Since we are interested in the potential relation between engagement and syntactic richness we decided to include modifiers and conjunctions frequency in our model even though the Pearson coefficient was small for both. Similarly, the fraction of time the speaker pauses during the turn was included in the feature set despite a coefficient equal to −0.08 (see Table 1).

3.5. Algorithms performance evaluation

We compare the performances of 7 different classifiers: *Logistic Regression, Support Vector Machines, K-Nearest Neighbors, AdaBoost, Naïve Bayes, Random Forest* and *Multilayer Perceptron*. Stratified 10-fold cross-validation is used to compute performance metrics for each algorithm. When relevant, hyperparameters are optimized following the nested cross-validation procedure (Scheffer, 1999). For each train-test fold, the training set is further divided between a training and a validation set used to estimate model performance for different combinations of parameters. The set of hyperparameters yielding the best results on the validation set was kept when evaluating the algorithm performance within the outer cross-validation loop. Hyperparameters taken into consideration for each classifier are indicated in Table 2. Regarding the Multilayer perceptron, the small size of our dataset encourages us to limit in practice the number of parameters. The architecture is therefore reduced to two hidden layers, the first with 64 nodes and the second with 32 nodes. Only the learning rate and the batch size were optimized using the nested cross-validation procedure. The number of epochs was set to 50.

Taking into account the fact that this study does not have direct equivalent in the literature (measuring engagement in naturalistic human-human conversation at the turn level in a 5-level scale), we have decided to experiment the classic models, in order to compare as directly as possible our results with the state of the art. We also designed two different baselines:

TABLE 2 Hyperparameters that were optimized using the nested cross-validation procedure for each classifier.

Model	Optimized hyperparameters
Logistic regression	None
SVM	Kernel: linear, polynomial, radial basis function Regularization: 0.1, 1, 10
K-nearest neighbors	Number of neighbors: 3, 5, 7, 9
AdaBoost	Number of estimators: 500, 1,000, 2,000
Naïve Bayes	None
Random forest	Number of estimators: 500, 1,000, 2,000 Maximum depth: 5, 10, None
Multilayer perceptron	Learning rate: 0.0001, 0.001, 0.01 Batch size: 16, 32

- Baseline 1: The classifier always predicts the majority class.
- Baseline 2: The classifier randomly selects labels according to their probabilities of occurrence in the whole dataset.

For each try, classifier permutation paired t-tests were conducted to assess statistical significance of the improvement over the baseline.

3.6. Modalities contributions

Given our interest in the importance of multimodality for engagement level classification, we adopted the procedure of Fedotov et al. (2018) in order to assess the respective strengths of modality contributions. Features were divided between prosodic-acoustic features (*log-energy, RMSE, f0*), prosodic-temporal features (*articulation rate, turn-taking delay, pauses*), mimo-gestural features (*smiles, head nods, arms and hands gestures*) and linguistic features (*modifiers and conjunctions frequency*). Classification was performed with each unimodal set separately as well as with every possible combination of two to three modalities. This operation was carried out using the classifier with the best weighted F-score.

4. Results

4.1. Descriptive statistics

4.1.1. Turns detection

Our 14 dyads corpus contains 1,486 automatically detected turns with duration greater than 1 second. Among them a total of 1,336 were manually annotated with engagement level, the remaining 150 turns being left out for three main reasons. First, moments when participants are not really conversing (e.g., reading

their short story or talking to experimenters), are sometimes wrongly identified as conversational turns. Second, even though turns shorter than 1 s are removed beforehand, some turns are still perceived as too short to allow any relevant annotation. Finally, turn boundaries are sometimes incorrectly identified because of a misalignment between tokens boundaries (as detected by SPPAS) and the audio signal. Finally, another case of incorrectly segmented turns, but that were annotated nevertheless, involves turns cut in two parts because the listener starts producing a short IPU while the main speaker is pausing without truly releasing the turn. The break is deemed erroneous insofar as the listener does not meaningfully take the turn. In this case the two parts were merged manually.

For a participant, the average number of annotated conversational turns is equal to 48.6. This number of turns is highly variable between participants, ranging from 20 to 97. Turn duration distribution is heavily skewed toward short turns. Most of them last for less than 5 s, while only a few last for more than 20 s, as shown in Figure 1. Turn duration does not seem to impact the level of engagement consistently, as demonstrated by the low Pearson correlation coefficient between turn duration and engagement level (see Table 1).

4.1.2. Annotation of engagement

As shown Figure 2, speech turns were predominantly labeled as either neutral (level 3) or engaged (level 4) with 41.6% and 28.2% of turns respectively (in a 5-levels scale). Speakers were rarely perceived as strongly disengaged. There is no significant difference in engagement level (Student test of $t = 1.49$, $p = 0.15$ on the mean engagement level) between the situation where participants know each other vs. first encounter.

4.2. Classification

The different experiments presented here vary in function of the number of classes. One of our goals in this paper is to see until what extent a fine-grained classification in 5 classes can be efficient, most of the works in the literature staying at a binary classification. We show the classification results obtained by different models.

4.2.1. 2 classes

Taking into account the distribution of the engagement annotations, we chose for reducing for 5 to 2 classes by merging levels 1 and 2 (disengaged) and levels 3, 4 and 5 (engaged). As reported in Table 3, results reach the state of the art with a F-score value at 0.78. Note that most of works in the literature relies on human-agent conversations, which is a slightly different task, features extracted from the agent's production being standardized.

4.2.2. 3 classes

As aforementioned subsection 4.1.2 strongly disengaged turns rarely occur in our dataset. This incites merging level 1 and 2 when classifying engagement degree. Other studies have adopted this strategy for similar reasons (Huang et al., 2016). Annotators also sometimes reported difficulty when deciding between attributing level 4 or level 5. We decided therefore to perform a new classification task, adopting that time 3 levels of engagement (i.e., level 1: strongly disengaged or disengaged, level 2: neutral and level 3: engaged or strongly engaged) rather than 5. Results of the 10-fold cross validation procedure are displayed in Table 4. The *SVM* achieves an F-score of 0.58, almost at the same performance as *Logistic Regression* and outperforming both the baseline and closest models (*Multilayer perceptron* and *KNN*) significantly.

4.2.3. 5 classes

Table 5 presents the weighted precision, recall and F-score obtained with each algorithm when classifying engagement level on the 5-level scale described above. All classifiers significantly outperform both baselines (see Section 3.5). The best results are obtained when using *KNN*. This classifier achieves a weighted F-score of 0.49. However performance was not found to be significantly different from the ones achieved by the *Logistic Regression* classifier.

4.3. Modality contribution

We used the best classifier in the 3-class situation for exploring modality contribution. Note at his stage that we wanted to explore the effect per modality, not per feature even though this second type of analyze, at a finer grain, is of course important. The question there is to try to have a higher-level type of information, maybe more easily explainable, trying to answer the specific role of each modality, whatever the features they cover. Classification performance was computed for each unimodal feature set (prosodic-acoustic, prosodic-temporal, mimo-gestural and morphosyntactic) and for each combination of two or three modalities, evaluating the 3-level engagement scale introduced Section 4.2.2 with the same stratified 10-fold cross-validation used above. Each box in Table 6 indicates the mean F-score obtained by training the two-layer perceptron on the corresponding unimodal or multimodal subset of features. For multimodal feature sets the statistical significance of the effect of adding a new modality to the previous set is indicated.

When training the multilayer perceptron on a single modality, the best results are obtained by using prosodic-acoustic and mimo-gestural features. As expected from the Pearson correlation coefficients, using the linguistic or the prosodic-temporal modalities alone lead to worse results. Moreover, combining any modality with linguistic features never leads to any significant improved performance. On the opposite, incorporating the prosodic-acoustic modality in the model almost always produces significantly better results, with the exception of the PA+PT combination. The mean F-score for the PA+MG is equal to 0.60, against 0.50 for MG only. The mean F-score for the PA+L combination is equal to 0.50, against 0.36 for linguistic feature set only. The best result is obtain with the PA+PT+MG combination (mean F-score = 0.61). Performance is not significantly different

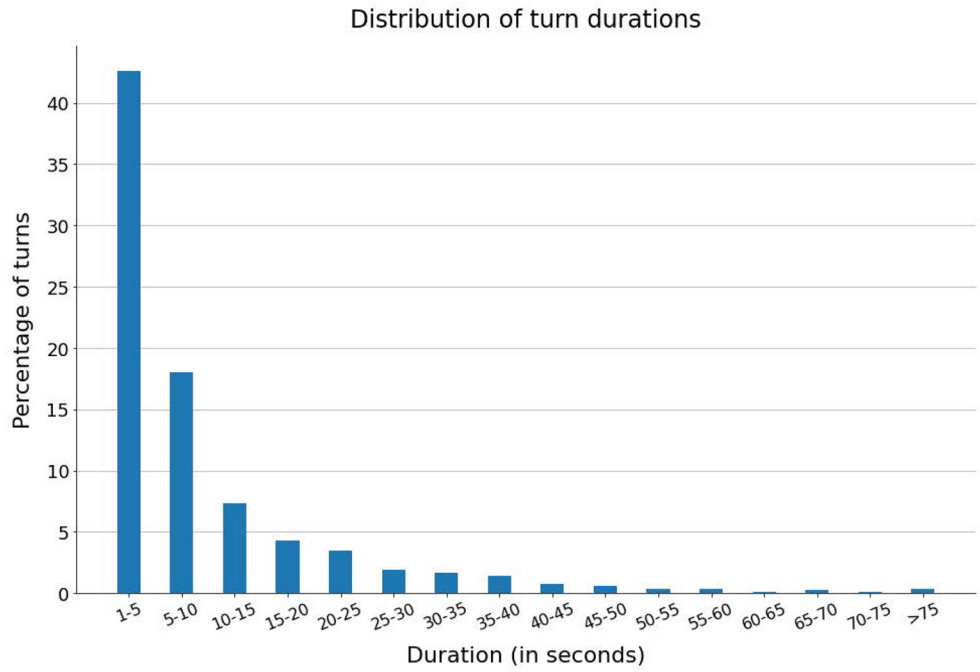

FIGURE 1
Distribution of turn durations (in seconds) in the dataset.

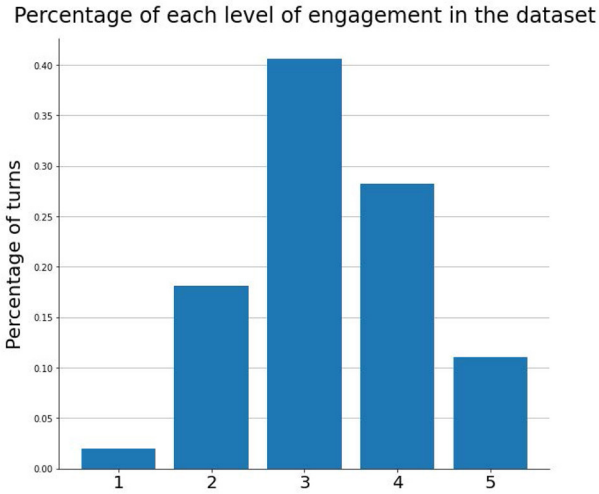

FIGURE 2
Distribution of engagement levels in the dataset.

from the one achieved with the PA+MG combination however, and merging all modalities yields similar results.

5. Discussion

All classifiers trained on our multimodal dataset significantly outperform the baselines in both the 5-class and the 3-class tasks. The best results were obtained with using logistic regression for the former, and with the two-layer perceptron for the latter. Limiting the number of classes to 3 by merging level 1 and 2 as well as level 4 and 5 leads to a small improvement in classification performance. Results for the 3-class task are comparable to the ones from previous studies. However it is worth noting that most of them were only interested in discriminating engagement from disengagement (Hsiao et al., 2012; Liu and Kappas, 2018; Ben-Youssef et al., 2019).

5.1. Characterizing engagement

Our results underline the importance of multimodality when classifying engagement levels. The best results are obtained by combining the prosodic-acoustic, prosodic-temporal and mimo-gestural modalities. An analogous performance is achieved when restricting the feature set to the combination of the prosodic-acoustic and mimo-gestural modalities. On the other hand, including morpho-syntactic features does not lead to any significant improvement in estimation performance. This suggests that engagement is not characterized by an increased complexity of the discourse, contrary to our hypothesis.

Our study also highlights the relevance of prosodic features when classifying engagement in human-human interactions. This contrasts with other studies conducted in analogous settings where this modality was either left out (Dermouche and Pelachaud, 2019) or found to not perform well compared to visual cues (Huang et al., 2016). On the other hand our results are consistent with findings from human-agent studies (Yu et al., 2004; Hsiao et al., 2012). Even though not as crucial as prosodic-acoustic features, mimo-gestural features were also found to be useful for engagement classification.

TABLE 3 Results of 2-class engagement level classification (standard deviation in parentheses).

Model	Precision	Recall	F-score	RMSE
Baseline 1	0.16 (±0.03)	0.40 (±0.04)	0.23 (±0.04)	0.77 (±0.00)
Baseline 2	0.66 (±0.04)	0.65 (±0.06)	0.66 (±0.05)	0.58 (±0.05)
Random forest	0.77 (±0.08)	0.80 (±0.01)	0.72 (±0.01)	0.44 (±0.01)
Logistic regression	0.77 (±0.03)	0.80 (±0.02)	0.77 (±0.02)	0.43 (±0.02)
SVM	0.63 (±0.004)	0.79 (±0.003)	0.70 (±0.003)	0.45 (±0.003)
K-nearest neighbors	0.74 (±0.03)	0.78 (±0.02)	0.75 (±0.02)	0.45 (±0.02)
AdaBoost	0.74 (±0.02)	0.71 (±0.06)	0.72 (±0.04)	0.52 (±0.06)
Naïve Bayes	0.78 (±0.02)	0.64 (±0.04)	0.67 (±0.03)	0.59 (±0.03)
Multilayer perceptron	0.77 (±0.02)	0.79 (±0.02)	**0.78** (±0.02)	0.45 (±0.02)

Best performances are highlighted in bold.

TABLE 4 Results of 3-class engagement level classification (standard deviation in parentheses).

Model	Recall	Precision	F-score	RMSE
Baseline 1	0.16 (±0.03)	0.40 (±0.04)	0.23 (±0.04)	0.77 (±0.00)
Baseline 2	0.33 (±0.04)	0.33 (±0.03)	0.33 (±0.04)	1.09 (±0.03)
Random forest	0.61 (±0.08)	0.59 (±0.04)	0.55 (±0.04)	0.71 (±0.03)
Logistic regression	0.59 (±0.05)	0.59 (±0.05)	0.58 (±0.05)	0.73 (±0.05)
SVM	0.61 (±0.04)	0.59 (±0.03)	**0.58** (±0.03)	0.70 (±0.05)
K-nearest neighbors	0.58 (±0.05)	0.57 (±0.05)	0.57 (±0.05)	0.73 (±0.06)
AdaBoost	0.54 (±0.03)	0.54 (±0.03)	0.53 (±0.03)	0.76 (±0.03)
Naïve Bayes	0.55 (±0.03)	0.52 (±0.03)	0.52 (±0.03)	0.82 (±0.05)
Multilayer perceptron	0.58 (±0.04)	0.57 (±0.04)	0.57 (±0.04)	0.78 (±0.05)

Best performances are highlighted in bold.

TABLE 5 Results of 5-class engagement level classification (standard deviation in parentheses).

Model	Precision	Recall	F-score	RMSE
Baseline 1	0.16 (±0.001)	0.40 (±0.001)	0.23 (±0.001)	0.98 (±0.004)
Baseline 2	0.31 (±0.03)	0.30 (±0.03)	0.30 (±0.03)	1.29 (±0.07)
Random forest	0.55 (±0.08)	0.52 (±0.02)	0.45 (±0.03)	0.81 (±0.03)
Logistic regression	0.49 (±0.02)	0.50 (±0.02)	0.48 (±0.02)	0.85 (±0.04)
SVM	0.48 (±0.06)	0.48 (±0.03)	0.43 (±0.06)	0.84 (±0.04)
K-nearest neighbors	0.50 (±0.03)	0.49 (±0.02)	**0.49** (±0.03)	0.88 (±0.08)
AdaBoost	0.35 (±0.04)	0.30 (±0.03)	0.29 (±0.03)	1.29 (±0.07)
Naïve Bayes	0.42 (±0.04)	0.35 (±0.04)	0.37 (±0.04)	1.23 (±0.06)
Multilayer perceptron	0.47 (±0.03)	0.48 (±0.02)	0.47 (±0.03)	0.87 (±0.07)

Best performances are highlighted in bold.

These results support the idea that the emotional dimension of engagement is expressed both through facial expressions and through prosody.

A last contribution of our study relates to interpretability. Indeed, training models with *AUs* models makes it difficult to later provide a meaningful characterization of engagement, even though some authors point out for example the relation between particular *AUs* and the expression of happiness (Ben-Youssef et al., 2019). The major limitation of attempting to characterize engagement using classifier such as the multilayer perceptron is that this does not allow assessing the direction of the effect each variable may have on the level of engagement. Still, inference can be done on the grounds of descriptive statistics. For example a greater quantity of laughing smiles (S4) seems to be associated with a higher degree of engagement, which is consistent with our initial hypotheses. Likewise, arms and hands gestures intensity, a feature rarely taken into account (Dermouche and Pelachaud, 2018), seems to be positively correlated with engagement. This later observation also supports the view of engagement as an investment in contributing to the conversation since co-speech gestures represent a physical effort toward making one's discourse more comprehensible. Greater gestures intensity could also simply

TABLE 6 Performance of unimodal and multimodal feature sets when classifying engagement level on a 3-level scale.

One modality	PA 0.52 (±0.03)	T 0.41 (±0.05)	MG 0.50 (±0.04)	V 0.36 (±0.04)		
Two modalities	PA-T 0.50 (±0.03)	PA-MG 0.60 (±0.02)	PA-V 0.50 (±0.04)	T-MG 0.52 (±0.04)	T-V 0.42 (±0.04)	MG-V 0.50 (±0.04)
Three modalities	PA-T-MG 0.61 (±0.02)	PA-T-V 0.51 (±0.03)	PA-MG-V 0.59 (±0.03)	T-MG-V 0.52 (±0.03)		
All modalities	PA-T-MG-V 0.60 (±0.03)					

For each weighted F-score standard deviation is indicated in parentheses (PA, prosodic-acoustic; T, temporal; MG, mimo-gestural; V, verbal).

signal increased excitation, which relates to the emotional aspect of engagement.

5.2. Looking at engagement at the turn level

Annotators' reports confirm the reliability of our turn detection algorithm: the small amount of incorrect turns were so because the participants were either reading the short story or talking with an experimenter. This supports its application to other corpora in order to increase the size of our dataset.

In the literature, turn-level engagement has only been investigated in the context of human-agent interactions (Forbes-Riley et al., 2012). The researchers achieved a F-score of 0.69 using a 2-level engagement scale. Our results from the 3-class classification task come close despite the more unconstrained nature of the interactions in our corpus and our more detailed engagement scale. The main limitation of such approach originates in the challenge of annotating heterogeneous turns with respect to engagement. Mental averaging is likely to have introduced additional noise in the data. Indeed, turns longer than 10 s, for which difficulties of annotation were often reported, represent approximately 40% of our dataset. Another limitation of focusing of turns relates to comparability, since the majority of studies are based on annotator-defined segments as already mentioned. It would be interesting to compare results when annotating engagement this way and when doing so on the basis of turns using the same dataset.

5.3. Future directions

Gaze direction was not included in our dataset. Given its central place in the engagement literature (Sidner et al., 2004; Ishii et al., 2021) it may be fruitful to incorporating it in future models. Including gaze direction may be far from being a silver bullet however, as some authors have found that it performs worse than facial features (Dermouche and Pelachaud, 2019). Head pose data might also be worthy of investigation given the good results obtained by Ooko et al. (2011) (0.89 mean F-score) when classifying engagement on a 3-level scale using a combination of head pose and head rotation. An additional interesting avenue would be to make use of AUs activation and intensity time series. The issue is that the framework used in previous studies using these features (Liu and Kappas, 2018; Ben-Youssef et al., 2019; Dermouche and Pelachaud, 2019) cannot be straightforwardly applied to our dataset to the extent that turns are of unequal sizes, which prevents from using an LSTM model with a fixed number of frames for example.

Another area worth of investigation relates to model building. As aforementioned, annotators have reported intra-turn variability in engagement level for turns longer than 10 s. In these cases annotation was carried out by mentally averaging variations in the level of engagement. One way to address this issue could be to design a multilevel model. Within each turn classification would be performed within sliding windows. Low-level predictions would then be aggregated based on either majority voting or on the statistical distribution of the predictions. The general idea is to reproduce the annotation process more accurately, especially when intra-turn variability is high. This model is actually being designed.

The aim of this paper was to propose a model for predicting the engagement level for each turn. The next step will be to analyze the dynamics of the engagement intra- and inter-participants. This means to study the engagement variations for each participant and identifies relative peaks of engagement for each of them. This also means to analyze the interplay between participants productions, how one participant behavior influence the other.

Compared to other works focusing on human-human interactions the size of our dataset (approximatively 1,200 turns after removing incorrect turns and turns for which feature values were missing), is fairly small. Increasing it might benefit the performance of our model. Another corpus of natural dyadic human-human conversations is currently being annotated. Once IPUs, feedbacks, smiles and head nodes are annotated, it will be possible to extract turns and then proceed to engagement annotation. At a smaller (and technical) scale, another way to alleviate this issue could be to use data augmentation techniques such as Smote as done in prior studies of engagement (Fedotov et al., 2018) in order to address the issue of imbalanced classes.

6. Conclusion

In this study we designed a multimodal model to classify engagement when participants assume the role of main speaker in talk-in interaction. For the 3-class case an F-score of 0.60 is reached using a simple two-layer perceptron. This encourages

exploring other avenues with regards to features extraction and selection, annotation procedure and model building. Our results demonstrate the importance of multimodality when automatically estimating engagement. Combining prosodic-acoustic and mimo-gestural features significantly improves classification performance compared to any unimodal feature set. This finding is also interesting insofar as prosodic information has not been the focal point of previous works focusing on human-human interactions. While engagement level classification at the turn level has already been performed in the context of human-agent interactions, this study is the first to adopt this approach when looking at human-human conversations. This method has the benefit of making it possible to study speaker's and listener's degree of engagement separately. When focusing on the former case, it allows to exploit a multimodal set of features. But additional challenges come with adopting this approach that favor a more fine grained analysis of the conversational structure, insofar as it requires detecting speech turns based on previous annotations of IPUs and feedbacks and sometimes makes annotating the level of engagement more complex. Aside from improving our current model the next step will consist in working on listeners' engagement.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Ortolang (https://hdl.handle.net/11403/paco/v1).

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AP-R conducted this study and mainly wrote the paper with PB. PB, RB, AB, and SR equally contributed to the work, that was supervised by PB. All authors contributed to the article and approved the submitted version.

Funding

This work was carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allwood, J., and Cerrato, L. (2003). "A study of gestural feedback expressions," in *First Nordic Symposium on Multimodal Communication* (Copenhagen), 7-22.

Amoyal, M., and Priego-Valverde, B. (2019). "Smiling for negotiating topic transitions in French conversation," in *GESPIN-Gesture and Speech in Interaction*.

Amoyal, M., Priego-Valverde, B., and Rauzy, S. (2020). "Paco: A corpus to analyze the impact of common ground in spontaneous face-to-face interaction," in *Language Resources and Evaluation Conference*.

Anzalone, S., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015). Evaluating the engagement with social robots. *Int. J. Soc. Robot.* 7, 465-478. doi: 10.1007/s12369-015-0298-7

Baker, R. S., Ocumpaugh, J., Gowda, S. M., Kamarainen, A. M., and Metcalf, S. J. (2014). "Extending log-based affect detection to a multi-user virtual environment for science," in *International Conference on User Modeling, Adaptation, and Personalization*. doi: 10.1007/978-3-319-08786-3_25

Bednarik, R., Eivazi, S., and Hradis, M. (2012). "Conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement," in *Workshop on Eye Gaze in Intelligent Human Machine Interaction*. doi: 10.1145/2401836.2401846

Ben-Youssef, A., Verni, G., Essid, S., and Clavel, C. (2019). On-the-fly detection of user engagement decrease in spontaneous human-robot interaction using recurrent and deep neural networks. *Int. J Soc. Robot.* 11, 815–828. doi: 10.1007/s12369-019-00591-2

Bickmore, T., Schulman, D., and Yin, L. (2010). Engagement in long-term interventions with relational agents. *Appl. Artif. Intell.* 24, 648–666. doi: 10.1080/08839514.2010.492259

Bigi, B. (2012). "Sppas: a tool for the phonetic segmentations of speech," in *The eighth international conference on Language Resources and Evaluation*, 1748–1755.

Blache, P., Abderrahmane, M., Rauzy, S., Ochs, M., and Oufaida, H. (2020). "Two-level classification for dialogue act recognition in task-oriented dialogues," in *Proceedings of COLING-2020*. doi: 10.18653/v1/2020.coling-main.431

Boersma, P., and Weenink, D. (1996). *Praat, a System for Doing Phonetics by Computer, Version 3.4*. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam.

Bohus, D., and Horvitz, E. (2009). "To predict engagement with a spoken dialog system in open-world settings," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. doi: 10.3115/1708376.1708411

Bonin, F., Bock, R., and Campbell, N. (2012). "How do we react to context? Annotation of individual and group engagement in a video corpus," in *Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom)*. doi: 10.1109/SocialCom-PASSAT.2012.110

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., and Blache, P. (2021a). "A multimodal model for predicting conversational feedbacks," in *International Conference on Text, Speech, and Dialogue*. Springer, 537-549. doi: 10.1007/978-3-030-83527-9_46

Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. (2009). "Detecting user engagement with a robot companion using task and social interaction-based

features," in *Proceedings of the International Conference on Multimodal Interfaces*. doi: 10.1145/1647314.1647336

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620539

Dermouche, S., and Pelachaud, C. (2018). "Analysis to modeling of engagement as sequences of multimodal behaviors," in *Language, Resources and Evaluation Conference (LREC)*.

Dermouche, S., and Pelachaud, C. (2019). "Engagement modeling in dyadic interaction," in *International Conference on Multimodal Interaction (ICMI '19)*. doi: 10.1145/3340555.3353765

Dhamija, S., and Boult, T. (2017). "Automated mood-aware engagement prediction," in *Seventh International Conference on Affective Computing and Intelligent Interaction*. doi: 10.1109/ACII.2017.8273571

Dybala, P., Ptaszynski, M., Rzepka, R., and Araki, K. (2009). Humans with humor : a dialogue system that users want to interact with. *IEICE Trans. Inf. Syst.* E92.D, 2394–2401. doi: 10.1587/transinf.E92.D.2394

Fedotov, D., Perepelkina, O., Kazimirova, E., Konstantinova, M., and Minker, W. (2018). "Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data," in *Workshop on Modeling Cognitive Processes from Multimodal Data*. doi: 10.1145/3279810.3279842

Forbes-Riley, K., Litman, D., Friedberg, H., and Drummond, J. (2012). "Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system," *Conference of the North American Chapter of the Association for Computational Linguistics*.

Foster, M., Gaschler, A., and Giuliani, M. (2017). Automatically classifying user engagement for dynamic multi-party human-robot interaction. *Int. J. Social Robot.* 9, 659–674. doi: 10.1007/s12369-017-0414-y

Glas, N., and Pelachaud, C. (2015a). "Definitions of engagement in human-agent interaction," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 944–949. doi: 10.1109/ACII.2015.7344688

Glas, N., and Pelachaud, C. (2015b). "Topic transition strategies for an information-giving agent," in *European Workshop on Natural Language Generation*. doi: 10.18653/v1/W15-4725

Gravano, A., and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* 25, 601–634. doi: 10.1016/j.csl.2010.10.003

Hsiao, C.-Y., Jih, W.-R., and Hsu, J. (2012). "Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns," in *Workshop on Activity Context Representation - Techniques and Languages (ACR12)*.

Huang, Y., Gilmartin, E., and Campbell, N. (2016). "Engagement recognition using auditory and visual cues," in *Interspeech 2016*. doi: 10.21437/Interspeech.2016-846

Ishii, R., Nakano, Y. I., and Nishida, T. (2013). Gaze awareness in conversational agents: estimating a user's conversational engagement from eye gaze. *ACM Trans. Interact. Intell. Syst.* 3, 249980 doi: 10.1145/2499474.2499480

Ishii, R., Ren, X., Muszynski, M., and Morency, L.-P. (2021). "Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling?" in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 131-138. doi: 10.1145/3472306.3478360

Khatri, C., Venkatesh, A., Hedayatnia, B., Gabriel, R., Ram, A., and Prasad, R. (2018). Alexa prize – state of the art in conversational ai. *AI Mag.* 39, e2810. doi: 10.1609/aimag.v39i3.2810

Leite, I., Martinho, C., and Paiva, A. (2013). Social robots for long-term interaction: a survey. *Int. J. Soc. Robot.* 5, 291–308 doi: 10.1007/s12369-013-0178-y

Leite, I., McCoy, M., Ullman, D., Salomons, N., and Scassellati, B. (2015). "Comparing models of disengagement in individual and group interactions," in *International Conference on Human-Robot Interaction (HRI)*. doi: 10.1145/2696454.2696466

Levinson, S. C., and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* 6. doi: 10.3389/fpsyg.2015.00731

Liu, T., and Kappas, A. (2018). "Engagement breakdown in hri using thin-slices of facial expressions," in *Thirty-Second AAAI Conference on Artificial Intelligence*.

Michalowski, M., Sabanovic, S., and Simmons, R. (2006). "A spatial model of engagement for a social robot" in *9th IEEE International Workshop on Advanced Motion Control, 2006*. 762–767. doi: 10.1109/AMC.2006.1631755

Mower, E., Feil-Seifer, D. J., Mataric, M. J., and Narayanan, S. (2007). "Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*. doi: 10.1109/ROMAN.2007.4415249

Nakano, Y. I., and Ishii, R. (2010). "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Conference on Intelligent User Interfaces (IUI)*. doi: 10.1145/1719970.1719990

Novielli, N. (2009). Hmm modeling of user engagement in advice-giving dialogues. *J. Multimodal User Interface* 3:131-140. doi: 10.1007/s12193-009-0026-4

Novielli, N., de Rosis, F., and Mazzotta, I. (2010). User attitude towards an embodied conversational agent: Effects of the interaction mode. *J. Pragm.* 42, 2385-2397. doi: 10.1016/j.pragma.2009.12.016

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., and Peters, C. (2020). Engagement in human-agent interaction: an overview. *Front. Robot. AI* 7, 92. doi: 10.3389/frobt.2020.00092

Oertel, C., De Looze, C., Scherer, S., Windmann, A., Wagner, P., and Campbell, N. (2011). "Towards the automatic detection of involvement in conversation," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*. doi: 10.1007/978-3-642-25775-9_16

Ooko, R., Ishii, R., and Nakano, Y. I. (2011). "Estimating a user's conversational engagement based on head pose information," in *10th International Conference on Intelligent Virtual Agents, IVA'11*. doi: 10.1007/978-3-642-23974-8_29

Peters, C., Castellano, G., and de Freitas, S. (2009). "An exploration of user engagement in HCI," in *International Workshop on Affective-Aware Virtual Agents and Social Robots*. doi: 10.1145/1655260.1655269

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005a). "A model of attention and interest using gaze behavior," in *Conference on Intelligent Virtual Agents (IVA)*. doi: 10.1007/11550617_20

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., Poggi, I., and Tre, U. R. (2005b). "Engagement capabilities for ECAS," in *AAMAS Workshop Creating Bonds with ACAs*.

Poggi, I. (2007). *Mind, hands, face and body: a goal and belief view of multimodal communication*. Wiedler.

Priego-Valverde, B., Bigi, B., and Amoyal, M. (2020). "Cheese!: a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor," *Language, Resources and Evaluation (LREC)*.

Rauzy, S., and Amoyal, M. (2020). "Smad: a tool for automatically annotating the smile intensity along a video record," in HRC2020, 10th Humour Research Conference.

Rauzy, S., Montcheuil, G., and Blache, P. (2014). "Marsatag, a tagger for french written texts and speech transcriptions," in Second Asia Pacific Corpus Linguistics Conference.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Scheffer, T. (1999). *Error estimation and model selection*. PhD thesis, UniversitätBerlin, School of Computer Science.

Sidner, C. L., and Dzikovska, M. (2002). "Human-robot interaction: Engagement between humans and robots for hosting activities," in *International Conference on Multimodal Interfaces*. doi: 10.1109/ICMI.2002.1166980

Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. (2004). "Where to look: a study of human-robot engagement," in *International Conference on Intelligent User Interfaces*. doi: 10.1145/964442.964458

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.* 166, 5. doi: 10.1016/j.artint.2005.03.005

Skarbez, R., Brooks, F. P., and Whitton, M. C. (2017). A survey of presence and related concepts. *ACM Comput. Surv.* 50, 3134301. doi: 10.1145/3134301

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inquiry* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Venkatesh, A. (2018). On evaluating and comparing open domain dialog systems. *arXiv: Comput. Lang.* doi: 10.48550/ARXIV.1801.03625

Witmer, B., and Singer, M. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence Teleoper. Virtual Environ.* 7, 225–240. doi: 10.1162/105474698565686

Yu, C., Aoki, P. M., and Woodruff, A. (2004). Detecting user engagement in everyday conversations. arXiv: arXiv [preprint] cs/0410027. doi: 10.21437/Interspeech.2004-327