The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions*

Angelo Cafaro
CNRS-ISIR, UPMC, France
cafaro@isir.upmc.fr

Johannes Wagner
Augsburg University, Germany
wagner@hcm-lab.de

Tobias Baur
Augsburg University, Germany
baur@hcm-lab.de

Soumia Dermouche
CNRS-ISIR, UPMC, France
dermouche@isir.upmc.fr

Mercedes Torres Torres
University of Nottingham, UK
pszmt1@nottingham.ac.uk

Catherine Pelachaud
CNRS-ISIR, UPMC, France
pelachaud@isir.upmc.fr

Elisabeth André
Augsburg University, Germany
andre@hcm-lab.de

Michel Valstar
University of Nottingham, UK
michel.valstar@nottingham.ac.uk

ABSTRACT

We present a novel multi-lingual database of natural dyadic novice-expert interactions, named NoXi, featuring screen-mediated dyadic human interactions in the context of information exchange and retrieval. NoXi is designed to provide spontaneous interactions with emphasis on adaptive behaviors and unexpected situations (e.g. conversational interruptions). A rich set of audio-visual data, as well as continuous and discrete annotations are publicly available through a web interface. Descriptors include low level social signals (e.g. gestures, smiles), functional descriptors (e.g. turn-taking, dialogue acts) and interaction descriptors (e.g. engagement, interest, and fluidity).

CCS CONCEPTS

•Information systems → Database design and models; *Semi-structured data; Data streams;* **•Human-centered computing → Systems and tools for interaction design;**

KEYWORDS

Affective computing, multimodal corpora, multimedia databases

ACM Reference format:
Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In *Proceedings of 19th ACM International Conference on Multimodal Interaction, Glasgow, Scotland, November 13-17, 2017 (ICMI'17),* 10 pages.
DOI: 10.475/123_4

*Produces the permission block, and copyright information

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in

ICMI'17, Glasgow, Scotland

© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06…$15.00
DOI: 10.475/123 4

1 INTRODUCTION

Natural and realistic human-human interactions are needed in order to support the analysis and increase understanding of human social behavior. The use of training materials based on those interactions is crucial for a variety of tasks, such as improving the automatic detection and interpretation of social signals, or to create computational models of realistic human social behavior. An interesting aspect of human social interaction is the occurrence of unexpected events, such as conversational interruptions, that impact the interactant's behaviors and cognitive state. This topic is understudied in the areas of Affective Computing [36] and Social Signal Processing [48].

This paper presents **NoXi**: the **NO**vice e**X**pert **I**nteraction database. NoXi is a database that is annotated and thoroughly designed for studying and understanding human social behavior during an information retrieval task targeting multiple languages, multiple topics, and the occurrence of unexpected situations. NoXi is a corpus of screen-mediated face-to-face interactions recorded at three locations (France, Germany and UK), spoken in seven languages (English, French, German, Spanish, Indonesian, Arabic and Italian) discussing a wide range of topics. Our first aim was to collect data from natural dyadic knowledge exchanges in the context of the H2020 project ARIA-VALUSPA [47] (*Artificial Retrieval of Information Assistants – Virtual Agents with Linguistic Understanding, Social skills and Personalized Aspects*).

NoXi has been designed from the onset with the aim of being used by a wider audience of researchers in a variety of applications other than information retrieval (c.f. Section 3). The dataset offers over 25 hours of recordings of dyadic interactions in natural settings, featuring synchronized audio, video, and depth data (using a Kinect 2.0). We aimed to obtain data of spontaneous behavior in a natural setting on a variety of discussion topics. Therefore, one of the main design goals was to match recorded participants based on their common interests. This means that we first gathered potential experts willing to share they knowledge about one or more topic they were knowledgeable and passionate about, and secondly we recruited novices willing to discuss or learn more about the available set of topics offered by experts.

Eliciting unexpected situations was another emphasis in creating NoXi, and therefore it includes controlled interruptions made by a confederate to one of the participants during the recordings as well as spontaneous interruptions made by one of the interactants in a subset of recordings. Both manual and automatic, discrete and continuous annotations are included in NoXi describing low level behavior (e.g. head movements, gestures, etc.) and high level user states such as arousal or interest, as well as, speech transcriptions on word and sentence level. A web interface is publicly available, thus allowing the research community to search the database using a variety of criteria (e.g. topic, language, etc...) and download the data. In this paper, NoXi's design and annotations are presented.

2 RELATED WORK

In recent years, a large number of multimodal datasets on human interaction have been collected and published [29, 38, 45, 46]. We briefly review some by dividing them into two groups, according to the number of interlocutors.

Dyadic interactions. In these datasets, interactions include two interlocutors, who are recorded carrying out both structured and unstructured conversations. There are numerous examples of such datasets, with a wide range of applications, such as speech recognition [15], behavior analysis [50], segmentation, emotion recognition [12] and depression detection [16]. Arguably, one of the most popular datasets of one-to-one interactions is SEMAINE [30]. A large audio-visual dataset created originally as part of an iterative approach to building virtual agents that can engage a person in a sustained and emotionally-colored conversation, the SEMAINE dataset was collected using the Sensitive Artificial Listener (SAL) paradigm. Following this paradigm, conversations were designed to include an abundance of nonverbal features, which both sustained conversation and signaled emotion. SEMAINE has been successfully applied to many computer vision problems, such as facial muscle action (FAC) detection [23], head nod and shake detection [21], non-verbal communication estimation [19], segmentation [33] and emotions [41].

Other relevant multi-modal datasets featuring dyadic interactions include: the Cardiff Conversation Dataset (CCDB) [3], an audio-visual database focusing on non-scripted interactions that do not predetermine the participants' roles (speaker/listener); the MAHNOB-Mimicry dataset, designed to analyze mimicry in dyadic scenarios where subjects act with a significant amount of resemblance and/or synchrony [43]; the SEWA project[1], which include video-chat recordings, audio transcript and hand-gesture annotations of human interactions; and the Interactive Emotional Dyadic Motion CAPture (IEMOCAP) dataset, which contains markers on the face, head, and hands, of interview participants during scripted and spontaneous spoken communication scenarios [10]. Furthermore, an interesting collection of multimodal datasets can be found at TalkBank [27], a web-accessible database of audio-visual recordings of both human and animal communication. One of the major benefits of TalkBank is that it also includes the transcripts of the conversations recorded.

Multi-party interactions. In these datasets, the number of interlocutors is not constrained. Popular examples of such datasets are the Belfast storytelling dataset [29], which collects spontaneous social interactions with laughter, the Multimodal multiperson corpus of laughter in interaction [34] which collects multimodal data of laughter with focus on full body movements and different laughter types, and the AMI Meeting Corpus [44], which collects multimodal data from recordings of meetings.

In NoXi we focused on dyadic interactions. However, contrary to previous work, we had an innovative setup that integrated mediated face-to-face interactions, i.e. participants interacted through a screen in different rooms, focused on information exchange and retrieval on a rich variety of topics, and it has been recorded in multiple languages (mainly English, French and German). Moreover, the dataset is grounded on a recording protocol thoroughly designed for emphasizing the emergence of unexpected situations during the interaction, such as conversational interruptions.

3 NOXI

The idea behind NoXi was to obtain a dataset of natural interactions between human dyads in an expert-novice knowledge sharing context. In a recording session one participant assumes the role of an expert and the other participant the role of a novice. When recruiting participants, potential experts offered to discuss about one or more topics they were passionate and knowledgeable about, whereas novices applied to a recording session based on their willingness to discuss, learn more and retrieve information about a topic of interest among those offered by experts. A matching of interests was found when a novice chose an expert's topic, then the dyad was recorded. This served the purpose of obtaining spontaneous dialogues on a variety of different topics for which the participants were passionate/knowledgeable about.

3.1 Design Principles

We prepared the recording protocol with the following design principles.

Setting. The schema in Figure 1 shows the physical setup. We opted for a screen-mediated recording for a twofold purpose, first it allowed us to record a face-to-face conversation without the need of multiple cameras recording from different angles as in classical face-to-face settings of other corpora (e.g. [29, 34]). Secondly, this setup is closer to a scenario where a virtual agent is displayed on a screen. Participants were recorded while standing to meet the second design principle listed as follows. **Data.** We aimed at capturing full body movements (e. g. postural changes, torso leaning) in addition to facial expressions, gestures and speech. Furthermore, we wanted to enrich the dataset with the above mentioned data and possibly new formats not being captured in existing databases (e. g. Kinect depth maps).

Interaction. We wanted to record spontaneous interactions but at the same time we were interested in (1) knowledge transfer, (2) information retrieval and (3) occurrences of unexpected events (e. g. interruptions). Participants were allowed to continue their conversation until it reached a natural end. This results in quite long interactions (for such a database): the minimum duration for a recording session was set to 10 minutes.

[1] https://sewaproject.eu

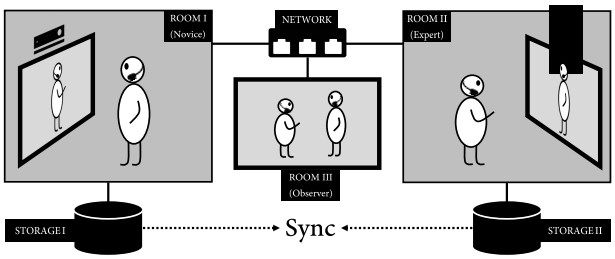

Figure 1: Sketch of the recording setup: Novice (left) and expert (right) are located in separate rooms while having a screen-mediated conversation. The interaction is monitored from a third room (middle) and recorded in sync.

Table 1: List of the recorded signals. Audio signals were sampled at 48 KHz, video signals at 25 fps.

Sensor	Channel	Resolution	Depth
Kinect	Audio	mono	16 bit signed
	Video	1920 x 1080	24 bit unsigned
	Depth	512 x 424	8 bit unsigned
	Skeleton	25 joints	32 bit float
	Confidence	25 values	32 bit float
	Face	1347 points	32 bit float
	Head	Pitch, roll, yaw	32 bit float
	Action Units	17 values	32 bit float
Headset	Audio	mono	16 bit signed

Participants. We primarily recruited participants in our research facilities, but also from our immediate social surroundings. Therefore we obtained dyads of both colleagues as well as entirely unacquainted persons, thus providing us the opportunity to record a number of zero-acquaintance situations [2].

Unexpected Events. One of the aims was to obtain occurrences of unexpected events. We primed both participants before recording a session by encouraging them to interrupt each other, provide opinions, suggest slight topic changes and induce a mild debate whenever possible. Moreover, we artificially injected an unexpected event during the recordings. More specifically, we introduced two possible functional interruptions (e.g. pretending that the microphone was not in the good position) that would result in an unexpected event, for the expert, from an external source (i.e. not being within the interaction or caused by the interactants). For this purpose, we informed the novice about the possibility to (1) call him on his/her mobile phone (i.e. CALL-IN) or (2) physically enter the recording room (i.e. WALK-IN). In either case we always interrupted the session at about 5 minutes after it began.

Recording Protocol The recording protocol had several steps. We first received participants in two different rooms. Figure 1 depicts the physical arrangement of the participants in the two rooms for the novice (Room I on the left side of the picture) and the expert (Room II on the right side). Room III (i.e. observer), depicted at the center of the picture, was used to monitor the session and synchronize the data collection. We primed the novice about the functional interruption as described above. In zero-acquaintance situations we did not introduce participants to each other, therefore their very first interaction happened when they both saw each other on the displays. We read instructions to both participants, set up their microphone and showed the position where to stand (also indicated by a marker on the floor) and prior to begin the recordings we obtained their informed consent.

Participants were informed about the sole possible usage of the recorded data for scientific research and non-commercial applications. Moreover, they had three (non exclusive) choices concerning the usage of their data: (1) data available within the ARIA-VALUSPA project's consortium only, (2) data available for dissemination purposes in academic conferences, publications and/or as part of teaching material, and (3) data available for academic and non-commercial applications to third-party users through the web interface described later in this section.

The session was monitored in Room III and when both participants agreed to end, the experimenter(s) gave participants questionnaires (described in Section 3.3. Finally, participants were debriefed and compensated.

3.2 Recording System

We used the Microsoft's Kinect 2 as recording devices. Kinect supports the capture of video streams in Full HD quality and provides optical motion capturing to track the body and face position of the user. The inbuilt microphone was used to capture the ambient sound in the room. Additionally, to obtain low-noise recordings of the voice we equipped users with a dynamic head-set microphone (Shure WH20XLR connected through a TASCAM US-322). The setup, as depicted in Figure 1, was distributed over three rooms. The rooms for the novice and the expert were equipped with a Kinect device put on top of a 55" flat screen. Kinect and headset are plugged to a PC (i7, 16 GB RAM). In each room a local hard drive (2 TB) was used to store the captured signal streams. A third PC was put in the observer room to monitor the interaction. The three PCs were connected in a wired local network.

The signals separately recorded for each user are listed in Table 1. Skeleton data had 14 values per joint, whereas for the face we had 3 values per point. The raw captured streams would require 154 MB of drive space per second for a single user. To ease the storage load two compromises were made: (i) the size of the HD video stream was reduced by applying the lossless Ut Video Codec by Takeshi Umezawa[2]. The algorithm builds on the Huffman code, but allows for a better compression. Since it runs on multiple threads and uses SSE2 assembly, it is fast enough to compress HD videos in real-time. (ii) the size of the depth images was reduced by decoding each pixel with only one instead of two bytes (i. e. depth values are expressed in the range of [0…255] instead of [0…60000]). The remaining streams (2 x audio, skeleton, confidence, face, head, and face animation units) were stored uncompressed. With the aid of these measures we were able to reduce the bandwidth from ~9.3 GB to ~1.4 GB per minute per user.

[2]http://umezawa.dyndns.info/archive/utvideo/

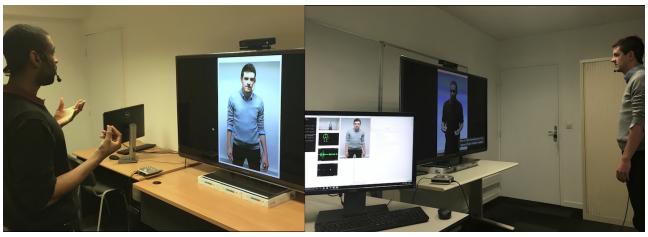

Figure 2: Snapshots of a novice-expert dyad in a recording session.

To stream audio-visual information from the novice room to the expert room and vice versa, as well as from the novice and expert room to the observer, we needed a very fast and efficient streaming protocol. Due to its popularity in streaming applications, we decided to use the h264 codec provided by the FFmpeg project[3]. For the sake of speed we also decided not to use a streaming server, but to stream directly to the receiver(s). Because of the way the experiment was designed we assumed that participants would stay more or less in the marked spot throughout the recording. Given the horizontal orientation of the video image, we decided to crop the streamed images from full HD to 480 × 720 pixels, thus discarding unused parts (i.e. left and right). Figure 2 shows an expert during interaction with a novice.

To keep recorded signals in sync we rely on a two-step synchronization. Once all sensors are properly connected and provide a stable data stream, we use a network broadcast to simultaneously start the recordings in the novice and expert room. In the following we then ensure that the captured streams keep a stable sample rate to avoid time drifts between the individual signals. The latter is achieved by regularly checking the number of received samples against the number of expected samples. In the cast that a discrepancy is observed either missing samples are added or additional samples are removed.

The described system was implemented with the Social Signal Interpretation SSI framework [49]. SSI is an open-source[4] framework for recording, analysing and fusing social signals in real-time.

3.3 Collected Data

The experiment was conducted in three different countries – France, Germany and UK. The primary reason for recording in three demographically different locations was the aim of collecting large numbers of interactions of the three languages targeted in the ARIA-VALUSPA project. In addition to English, French, and German, we also collected a smaller number of recordings of four other languages (Spanish, Indonesian, Italian, Arabic). A summary of the recorded sessions is given in Table 2. For the three main languages English, French, and German, we had 40, 25, and 19 interactions. In total, 87 people were recorded during 84 dyadic interactions (some people appeared in more than one session). The total duration of all sessions was 25 hours and 18 minutes. For each participant (i.e. expert or novice) in a given session, Table 3 shows the data that has been collected during the sessions and that is publicly available for download, in addition to the annotations described in Section 4.

[3]https://ffmpeg.org/
[4]http://openssi.net

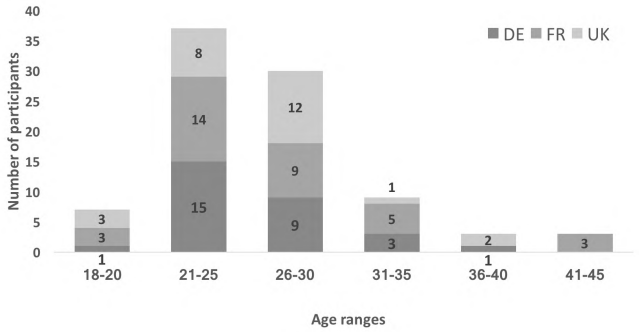

Figure 3: Age distribution of participants recorded in the three locations: Germany (DE), France (FR) and United Kingdom (UK).

Table 2: Overview of NOXI recordings - Sessions. From left to right: place, number of recording sessions, number of participants (female/male), average and standard deviation of recording duration (mm:ss), total duration (hh:mm).

Place	Sessions	Participants	Avg Dur	Std Dur	Total Dur
DE	19	29 (05/24)	17:56	05:56	05:38
FR	25	32 (10/22)	20:15	06:51	08:26
UK	40	26 (11/15)	16:50	06:41	11:13
Total	84	87 (26/61)	18:06	06:28	25:18

We also collected demographic information of our participants at the end of each session. These data consisted of their gender, cultural identity, age and education level. The cultural identity was obtained by asking participants to select the country that most represented their cultural identity from a list of all countries in the world. Participants' age is in the range of 21-50 years old. A breakdown of participants per age and recording location is shown in Figure 3.

In addition to demographic information, participants provided a self-assessment of their personality based on the Big 5 model (a.k.a. OCEAN) [28] by using the Saucier's Mini-Markers set of adjectives [40].

Finally, we collected session specific information that included the social relationship level between participants (e.g. acquaintances, friends), the level of expertise on the discussed topic and the proficiency level of the language spoken for that session.

We were very pleasantly surprised by the large diversity of topics covered by our experts. A total of 58 topics were discussed, the diagram in Figure 4 illustrates the most recurrent ones in the three main languages of the database. English sessions had a large variety of topics including travels (5 sessions), technology (4), health (3), cooking (3), sports (2), politics (2) and many others. French sessions were mainly about video-games (4), travels (3), music (3) and photography (2). Finally, German sessions included expert

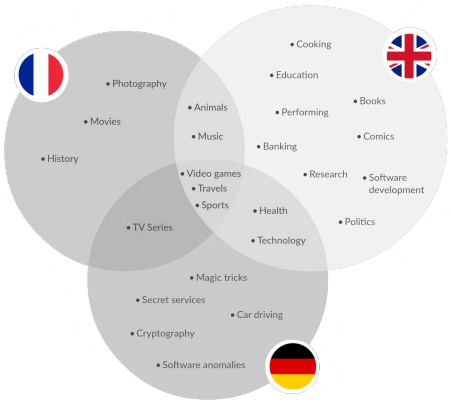

Figure 4: The most recurrent topics in the three main languages of NoXi.

computer science fields (5), various sports (6), car driving, magic tricks and other trivia.

3.4 Availability

NoXi is made freely available to the research community and for non-commercial uses. It is available through a web interface at: https://noxi.aria-agent.eu/. In the web interface, the data is organized in sessions that correspond to an expert-novice full recording with audio-visual data and annotations. The total size of the database is approximately 4 TB, however the database is searchable through the web interface. A user can select multiple criteria (e.g. language or topic of the session, participants' gender, etc...) and can choose the sessions to download from a list or results.

Furthermore, after selecting the sessions from the search results, a user can choose the type of files that s/he wishes to include in the download for the expert and the novice. Search results can be saved and shared with other NoXi users of the web interface. This is implemented through the notion of collections, which are predefined sets of sessions grouped by one or more criteria. Users can create a collection starting from the search results.

Table 3: List of recorded files available for download (* is replaced by Novice or Expert).

Sensor	Filename	Signal
Kinect	_kinect.wav	Audio (room)
	_video.avi	HD Video
	_video.mp4	Compressed Video
	_depth.stream	Depth Image
	_skel.stream	Skeleton Data
	_skelconf.stream	Skeleton Confidence
	_face.stream	Facial Points
	_head.stream	Head Position
	_au.stream	Action Units
Headset	_close.wav	Audio (close-talk)

4 ANNOTATIONS

Since the recording of NoXi, considerable efforts have been spent to collect descriptions of the database. Due to the large amount of data we wanted to allow multiple annotators from several sites to contribute and share their labeling efforts in an easy and clear way. In addition and to further speed up the process, we decided make use of semi-automated solutions to accomplish the desired descriptions. In the following section we will first introduce a novel annotation tool named NOVA implemented for this purpose. Afterwards, we report on descriptions already accomplished and describe in detail the annotation of conversational interruptions.

4.1 NOVA

Following the data collection, we were in need of a graphical tool to review and annotate the recorded data. Since the great majority of multi-modal corpora is limited to audio-visual data, there is a lack of tools with support for other type of signals, such as body or face tracking data. Hence, a database like NoXi can be viewed in parts only. The same is true if one aims to describe phenomena using different types of annotation schemes. For example, we might prefer a discrete annotation scheme to label behavior that can be classified into a set of categories (e.g. head nods and shakes), whereas a continuous scale would be more appropriate for variable dimensions like activation and evaluation. Since conventional tools only support either of the two schemes the annotation process is further complicated.

ARIA-VALUSPA aims at developing a novel graphical tool that overcomes the mentioned limitations and supports the visualization and description of arbitrary data streams using various annotation schemes. The tool is released under open-source license[5]. Figure 5 shows one of the NoXi sessions displayed with NOVA. It features audio-visual, as well as, skeleton and face streams of both participants. Below the signals several annotation tracks are shown including a speech transcription, a categorical description of voice activity and filler words, and valence/arousal scores expressed on a dimensional scale.

Moreover, NOVA offers a truly collaborative annotation platform that allows annotators from different sites to combine their efforts and immediately share the progress with each other. This has been realized by adding a database back-end, where annotations are stored and can be accessed together with the annotated signals. To create a new annotation session a user connects to a MongoDB server [6] where they can examine what kind of descriptions have been accomplished for the database so far. He or she then selects one of the pre-defined schemes and the information that should be displayed during the annotation process. The tool then creates an empty track and downloads selected media files. A set of security policies prevents situations where a user accidentally overwrites the annotation of another user. Only users with administration rights can edit and delete annotations of other users. This way an admin user can divide up forthcoming annotation tasks among the pool of annotators. Once a user has finished a task he or she can mark this in the database to signal availability for new tasks. Tools

[5] https://github.com/hcmlab/nova
[6] https://www.mongodb.com/

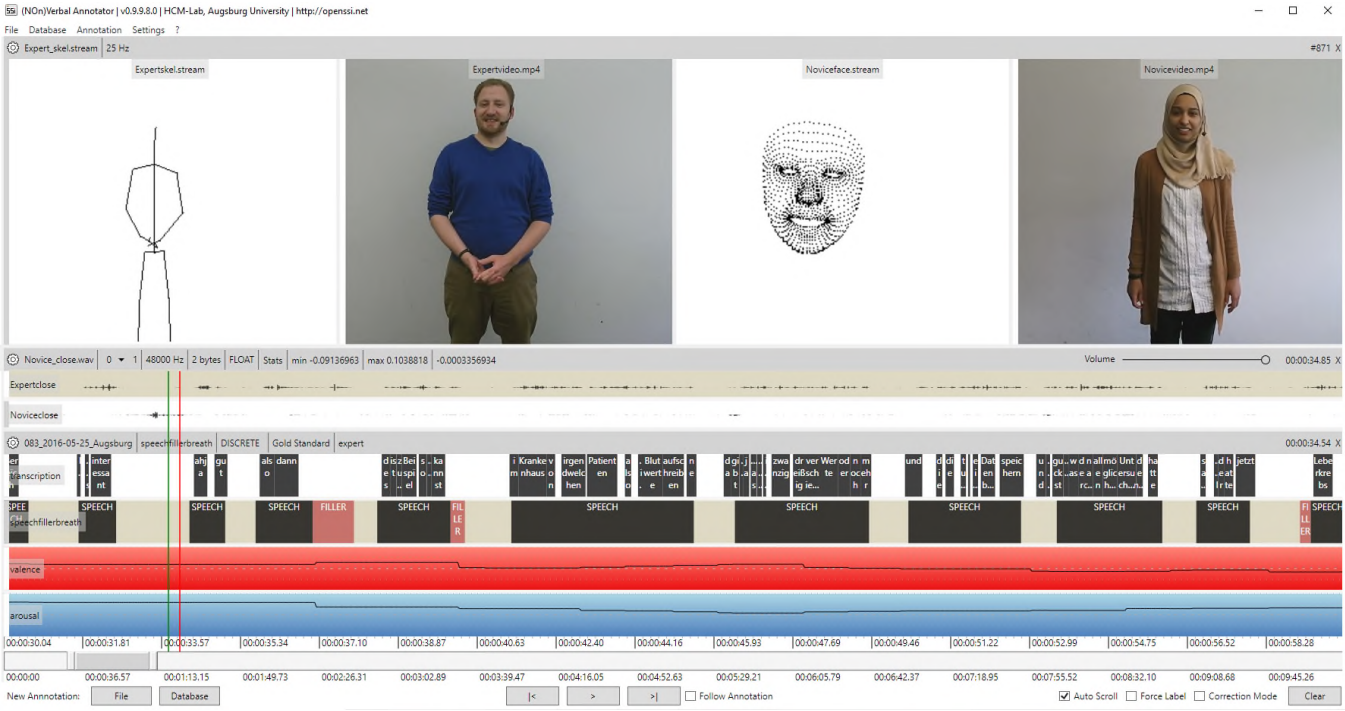

Figure 5: NOVA: On top videos of two users interacting during a recording and outputs of skeleton and face tracking. Audio streams are displayed as waveforms. At the bottom discrete and continuous annotations tracks are shown (see subsection 4.2).

are provided to merge the annotations from several users to create a gold standard.

However, the database back-end is not only meant to share efforts between human operators: it can also function as a mediator between human and machine. The idea is to minimize human labeling effort by letting the machine complete a task once sufficient information is available to train a machine learning model. To this end, two strategies are available from the NOVA interface:

(1) Session Completion: The annotator selects an annotation he would like to complete. The partly finished annotation is used to train a model and predict the remaining sessions. Optionally, predictions with a low confidence are highlighted.
(2) Session Transfer: The annotator selects a number of completed sessions and chooses the sessions he wants the machine to finish. The labeled sessions are then used to train a classification model and predict the target sessions. Again, predictions with a low confidence can be marked.

4.2 Available Descriptions

More than 30 annotators from three countries (UK, France, Germany) are involved in the annotation of NoXi. They use NOVA (see previous section) to create and share their annotations. To reduce human efforts where possible, we also apply automated or semi-automated methods.

Since participants were equipped with close-talk microphones only little background noise can be observed in the voice recordings. Hence, voice activity detection (VAD) was implemented by first normalizing the waveforms and afterwards applying a threshold to the intensity. Comparing the sequence of speech segments of both interlocutors makes possible studying their turn taking and interruption strategies (see next section for a detailed discussion). However, completely ignoring the semantic context of speech can lead to wrong annotations. For instance, overlapping segments (i.e. where both interlocutors talk simultaneously) may not necessarily signal an attempt to take the floor but can be a sign of backchanneling, too. Hence, VAD annotations were refined by marking filler events such as hesitation (e.g. "uhm") and backchanneling events (e.g. "ok"). Once a sufficient number of annotations had been manually processed, cooperative learning [17] was used to predict the annotations for the remaining sessions.

Skeleton tracking from the Kinect 2 was used to automatically assess gestures and movement quality. Gesture recognition was accomplished with the full body interaction framework FUBI [24]. It supports the recognition of static postures and dynamic gestures by comparing each skeleton frame against a series of recognition automata. To define the automata, FUBI offers a user-friendly XML based language. From the body we extracted gestures such as arm crossing or leaning front/back; from the head gestures such as nods and shakes. To capture the dynamic properties of the movements we calculated several expressiveness measures such as energy, fluidity, or spatial extend [5].

The emotional state of the user is currently described along three affective dimensions: valence, arousal and interest [25]. Each dimension is represented by a continuous score between 0 and

Table 4: Annotation scheme for the multimodal behaviors and engagement annotations in NoXi.

Tier (modality)	Labels
Audio	PAUSE_DURATION — PITCH — SIGNAL_INTENSITY
Head movements	NOD — SHAKE
Head direction	FORWARD — BACK — UPWARDS — DOWNWARDS — SIDEWAYS — SIDE_TILT
Smiles	SMILE
Eyebrow movements	RAISED — FROWN
Gaze direction	TOWARDS_INTERLOCUTOR — UP — DOWN — SIDEWAYS — OTHER
Gestures	ICONIC — METAPHORIC — DEICTIC — BEAT — ADAPTOR
Hand rest positions	ARMS_CROSSED — HANDS_TOGETHER — HANDS_IN_POCKETS — HANDS_BEHIND_BACK — AKIMBO
Engagement	STRONGLY DISENGAGED ... STRONGLY ENGAGED

1, which we automatically derive from the user's voice. To train the speech models we relied on the "Geneva Multimodal Emotion Portrayals" (GEMEP) corpus [4]. It contains 1.2 k instances of emotional speech from ten professional actors, which we used to train Support Vector Machines classifiers using the popular LibSVM library [14]. As feature set we took the ComParE 2013 set [42], which has been extensively demonstrated to be suitable for a wide range of paralinguistic tasks. A total of 6373 features were extracted on a per-chunk level using the OpenSMILE toolkit [18].

Finally, manual transcriptions on word and sentence level are available for different languages. These transcriptions help fine-tune the acoustic and language models, as well as, estimate the performance of the Automatic Speech Recognition (ASR) system implemented in ARIA-VALUSPA.

4.3 Multimodal Signals of Engagement

NoXi aims to provide data supporting research in embodied conversational agents (ECAs) for making them capable of interpreting users' non-verbal behaviors and prosody in real-time and associate it with different engagement variations (increase or decrease). This allows the agent to adapt to the user's behavior in order to maintain the desired level of engagement during the interaction. Thus, we describe an annotation schema for both experts and novices about engagement variation, non-verbal behaviors and prosody.

Several coding schemes have been used for multimodal corpus annotation. An exhaustive one is the MUMIN multimodal coding scheme [1] that we adapt for NoXi as summarized in Table 4. In addition to NOVA, we rely on Praat [8] for prosodic and acoustic features annotations.

Audio modality. We extracted several prosodic and acoustic features such as *pitch* (i.e. the quality of a sound represented by the rate of vibrations), pause duration (i.e. the duration of silence time in an audio segment) and signal intensity. In order to avoid content biases from the utterances when subjectively annotating engagement, we have filtered it out, for both interactants, by applying a Pass Hann Band Filter. With this method the voice fragments kept their prosodic information but lost their content.

Facial behavior. Several signals are semi-automatically annotated: gaze direction, head movements/direction, smile, and eyebrows movements.

Gestures. Several types of gestures are annotated based on McNeill's classification [31]: iconics, metaphorics, deictics, beats, and adaptors. We also include hand rest positions such as: arms crossed, hands together, hands in pockets, hands behind back and akimbo.

Engagement. A range of different definitions of engagement exist in human-agent interaction. For example, Poggi formulates engagement as: *"The value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction"* [37]. This definition has been used in several works [6, 13, 20, 35, 39]. Based on this definition we added continuous annotations of engagement and then provided discrete instances when a change occurs ranging from strongly disengaged to strongly engaged.

4.4 Conversational Interruptions

Another NoXi's purpose is to provide a dataset for analyzing and studying humans reactions to unexpected situations during the interaction. Conversational interruptions represent one of those situations. In face-to-face conversations, interruptions can be considered as turn-taking violations [7] (e.g., claiming the turn by interrupting the current speaker), but they can also serve as important social displays that reflect interpersonal attitudes (e.g., dominance or cooperation) as well as involvement in the interaction [32].

Table 5: Automatic annotations describing the conversation state and turn transitions for Expert and Novice in NoXi.

Tier	Labels
Conversation State	NONE EXPERT — NOVICE BOTH
Turn Transition	PAUSE_W — PAUSE_B PERFECT OVERLAP_W — OVERLAP_B

As a first step towards this direction we automatically analyzed NoXi's audio recordings and, based on both Expert and Novice's voice activity detection (VAD, see previous section), we automatically imply two annotation tiers: **communicative state** and **turn**

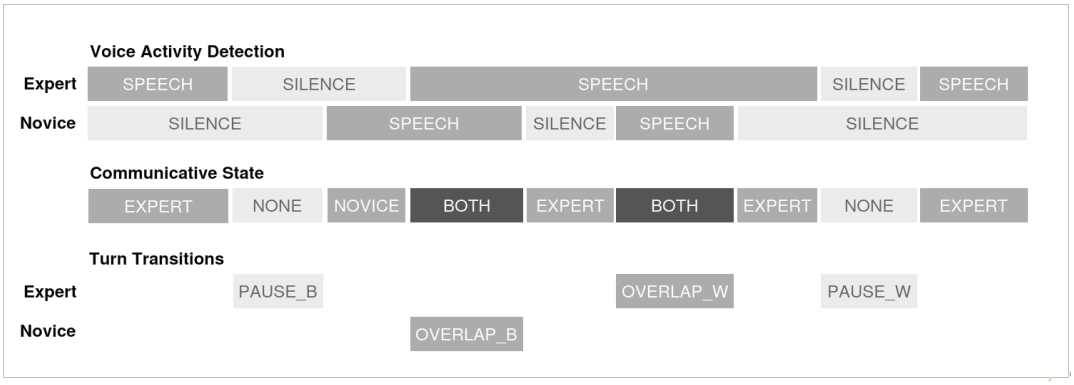

Figure 6: Example annotations of communicative states and turn transitions tiers automatically implied from Expert and Novice's voice activity detection.

transitions. Table 5 shows a summary of the labels for both annotation tiers. The communicative state tier describes both interlocutors' activity in the conversation, thus resulting in four states (i.e. annotation labels) describing when no one speaks (*NONE*), the expert speaks (*EXPERT*), the novice speaks (*NOVICE*) or both interlocutors speak (*BOTH*). For obtaining those labels, we ran the voice activity detection in off-line mode on the whole NoXi database. In particular, for each session we analyzed the individual recorded audio streams coming from the close talk microphone. We implemented a finite state machine that attributed the labels (NONE, NOVICE, EXPERT and BOTH) crossing the information about the voice activity that was detected.

For the turn transitions we followed the ideas from [22]. In the annotation schema, a turn transition is a change of the communicative state during the conversation that can result in a switch from speech to silence and vice-versa for the same speaker (i.e. within turn) or between the two speakers (i.e. between turns). Therefore, the following turn transitions are annotated in two separate tiers for Expert and Novice in a given session. A **pause within** turns (*PAUSE_W*) is a silence between two turns of the same speaker without speakers switch. A **pause between** turns (*PAUSE_B*) is a speaker switch from Expert to Novice (or vice-versa) with a silence in between. A **perfect** turn transition (*PERFECT*) happens with a speaker's change without a silence nor an overlap in between. An **overlap between** turns (*OVERLAP_B*) is marked when an overlap between speakers with a speaker change occurs. Whereas an **overlap within** a turn (*OVERLAP_W*) is a speakers' overlap without any speaker change.

In order to obtain those labels, we ran a parallel finite state machine that used state transitions of the former one (e.g. transition from EXPERT speaking to SILENCE and then EXPERT or NOVICE speaking) for labeling the turn transitions. Therefore, a conversation state that goes from EXPERT to SILENCE and then to EXPERT again, triggers a PAUSE_W (pause within turns) label in the turn transitions. Figure 6 shows an example annotation of these two tiers. In the turn transitions tiers, the labels are always assigned to the tier corresponding to the current speaker. The first pause between turns (*PAUSE_B*), for instance, is assigned to the Expert tier because there is a transition that goes from the expert to the novice through a pause in the expert's speech. The novice's tier obtains an overlap between turns because while the novice is speaking both start speaking and then the turn goes to the expert.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented NoXi, a novel multi-lingual database of natural dyadic expert-novice interactions focused on unexpected situations (e.g. conversational interruptions) and adaptive behavior (e.g. engagement). NoXi offers a rich set of audio-visual synchronized data recorded in screen-mediated interactions in the context of information exchange and retrieval. Audio-visual continuous and discrete descriptors are automatically and semi-automatically provided. Those range from low level social signals such as gestures and smiles, to higher level descriptors of interest and engagement. A web interface makes the whole set of NoXi's data and annotations publicly available for non-commercial purposes.

In the short term we plan to add new annotations that will be available through the web interface. In particular, we aim at providing discrete annotations of dialogue acts (or dialogue moves) based on the Dynamic Interpretation Theory (DIT++) taxonomy, a comprehensive application-independent ISO standard [9], for the classification and analyses of dialogue with information about the communicative acts that are performed by dialogue segments.

As for conversational interruptions, future work includes the higher level annotation of the strategy employed by the interrupter when interrupting the current speaker [11]. Therefore, we aim at distinguishing disruptive and cooperative interruptions, respectively interruptions that are intended to help the speaker by coordinating the process and/or content of the ongoing conversation [26], as opposed to intrusive interruptions that pose threats to the current speaker's territory by disrupting the process and/or content of the ongoing conversation [32]. Finally, we plan to exploit NoXi's data and annotations for modeling and generating agent's verbal and nonverbal behavior in the context of ARIA-VALUSPA.

ACKNOWLEDGMENTS

This work is funded by European Union Horizon 2020 research and innovation programme, grant agreement No 645378.

REFERENCES

[1] J Allwood, L.Cerrato, K Jokinen, C.Navarretta, and P.Paggio. 2007. The MUMIN annotation scheme for feedback, turn management and sequencing. *International Journal of Language Resources and Evaluation* (2007), 1–18. DOI: https://doi.org/10.1007/s10579-007-9061-5

[2] Nalini Ambady, Mark Hallahan, and Robert Rosenthal. 1995. On Judging and Being Judged Accurately in Zero-Acquaintance Situations. *Journal of Personality and Social Psychology* 69, 3 (1995), 518–529.

[3] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. 2013. Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 277–282.

[4] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. 2012. Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception. *Emotion* 12, 5 (2012), 1161–1179.

[5] Tobias Baur, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, and Patrick Gebhard. 2015. Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 2 (2015), 11.

[6] Tobias Baur, Dominik Schiller, and Elisabeth André. 2016. Modeling Users Social Attitude in a Conversational System. In *Emotions and Personality in Personalized Services*. Springer International Publishing, 181–199.

[7] Geoffrey W Beattie. 1981. Interruption in conversational interaction, and its relation to the sex and status of the interactants*. *Linguistics* 19, 1-2 (1981), 15–36.

[8] Paul Boersma. 2001. Praat, a System for Doing Phonetics by Computer. *Glot International* 5, November 2000 (2001), 341–345.

[9] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation.. In *LREC*. 430–437.

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[11] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *Proc. of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS fi16)*. International Foundation for Autonomous Agents and Multiagent Systems, 911fi?!920.

[12] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 375–388.

[13] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. *Proc. of the 2009 international conference on Multimodal interfaces* January 2009 (2009), 119. DOI: https://doi.org/10.1145/1647314.1647336

[14] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages. DOI: https://doi.org/10.1145/1961189.1961199

[15] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[16] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F Cohn. 2017. Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics* (2017).

[17] Miaobo Dong and Zengqi Sun. 2003. On human machine cooperative learning control. In *Proc. of the 2003 IEEE International Symposium on Intelligent Control*. 81–86.

[18] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proc. of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 835–838. DOI: https://doi.org/10.1145/2502081.2502224

[19] Florian Eyben, Martin Wöllmer, Michel F Valstar, Hatice Gunes, Björn Schuller, and Maja Pantic. 2011. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proc. of IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 322–329.

[20] Nadine Glas and Catherine Pelachaud. 2014. Politeness versus Perceived Engagement: an Experimental Study. In *The 11th International Workshop on Natural Language Processing and Cognitive Science*. Venice, Italy.

[21] Hatice Gunes and Maja Pantic. 2010. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *International conference on intelligent virtual agents*. Springer, 371–377.

[22] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 4 (2010), 555–568. DOI: https://doi.org/10.1016/j.wocn.2010.08.002

[23] Bihan Jiang, Michel F Valstar, and Maja Pantic. 2011. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 314–321.

[24] Felix Kistler and Elisabeth André. 2015. How Can I Interact? Comparing Full Body Gesture Visualizations. In *Proc. of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '15)*. ACM, New York, NY, USA, 583–588. DOI: https://doi.org/10.1145/2793107.2810299

[25] Shinobu Kitayama, Hazel Rose Markus, and Masaru Kurokawa. 2000. Culture, emotion, and well-being: Good feelings in Japan and the United States. *Cognition & Emotion* 14, 1 (2000), 93–124.

[26] Han Z Li. 2001. Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse. *Journal of Language and Social Psychology* 20, 3 (2001), 259–284.

[27] Brian MacWhinney. 2007. The TalkBank Project. In *Creating and digitizing language corpora*. Springer, 163–180.

[28] Robert R. McCrae and Jr. Costa, Paul T. 1997. Personality trait structure as a human universal. *American Psychologist* 52, 5 (1997), 509–516. DOI: https://doi.org/10.1037/0003-066X.52.5.509

[29] Gary McKeown, William Curran, Johannes Wagner, Florian Lingenfelser, and Elisabeth André. 2015. The Belfast Storytelling Database – A spontaneous social interaction database with laughter focused annotation. In *International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*.

[30] Garry McKeown, Michel F. Valstar, Roddy Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *International Conference on Multimedia and Expo (ICME)*. 1079–1084.

[31] David McNeill. 1992. *Hand and mind : what gestures reveal about thought*. University of Chicago Press, Chicago. 416 pages.

[32] Kumiko Murata. 1994. Intrusive or co-operative? A cross-cultural study of interruption. *Journal of Pragmatics* 21, 4 (1994), 385–400.

[33] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2010. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. (2010).

[34] Radoslaw Niewiadomski, Maurizio Mancini, Tobias Baur, Giovanna Varni, Harry Griffin, and Min SH Aung. 2013. MMLI: Multimodal multiperson corpus of laughter in interaction. In *International Workshop on Human Behavior Understanding*. Springer International Publishing, 184–195.

[35] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. Engagement Capabilities for ECAs. In *AAMAS'05 workshop on Creating Bonds with ECAs*.

[36] Rosalind W Picard and Roalind Picard. 1997. *Affective computing*. Vol. 252. MIT press Cambridge.

[37] I. Poggi. 2007. *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*.

[38] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–8.

[39] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proc. of the 6th international conference on Human-robot interaction*. 305. DOI: https://doi.org/10.1145/1957656.1957781

[40] Gerard Saucier. 1994. Mini-Markers: A Brief Version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment* 63, 3 (1994), 506. DOI: https://doi.org/10.1207/s15327752jpa6303_8

[41] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Bjorn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wollmer. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Trans. Affect. Comput.* 3, 2 (April 2012), 165–183. DOI: https://doi.org/10.1109/T-AFFC.2011.34

[42] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.. In *Interspeech: 14th Annual Conference of the International Speech Communication Association*, Frdric Bimbot, Christophe Cerisara, Ccile Fougeron, Guillaume Gravier, Lori Lamel, Franois Pellegrino, and Pascal Perrier (Eds.). ISCA, 148–152.

[43] X. Sun, J. Lichtenauer, M. F. Valstar, A. Nijholt, and M. Pantic. 2011. A Multimodal Database for Mimicry Analysis. In *Proc. of the 4th Bi-Annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*. Memphis, Tennessee, USA.

[44] Fabio Valente, Samuel Kim, and Petr Motlicek. 2012. Annotation and Recognition of Personality Traits in Spoken Conversations from the AMI Meetings Corpus.. In *INTERSPEECH*. 1183–1186.

[45] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

[46] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. 2015. Fera 2015-second facial expression recognition and analysis challenge. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015*, Vol. 6. IEEE, 1–8.

[47] Michel Valstar et al. 2016. Ask Alice: An Artificial Retrieval of Information Agent. In *Proc. of the 18th ACM International Conference on Multimodal Interaction*. ACM, 419–420.

[48] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.

[49] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proc. of the 21st ACM international conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 831–834.

[50] Zhihong Zeng, Jilin Tu, Ming Liu, Thomas S Huang, Brian Pianfetti, Dan Roth, and Stephen Levinson. 2007. Audio-visual affect recognition. *IEEE Transactions on multimedia* 9, 2 (2007), 424–428.