# Audience Engagement Prediction in Guided Tours through Multimodal Features

**Andrea Amelio Ravelli, Andrea Cimino, Felice Dell'Orletta**
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{andreaamelio.ravelli,andrea.cimino,
felice.dellorletta}@ilc.cnr.it

## Abstract

This paper explores the possibility to predict audience engagement, measured in terms of visible attention, in the context of guided tours. We built a dataset composed of Italian sentences derived from the speech of an expert guide leading visitors in cultural sites, enriched with multimodal features, and labelled on the basis of the perceivable engagement of the audience. We run experiments in various classification scenarios and observed the impact of modality-specific features on the classifiers.

## 1 Introduction

During face-to-face interactions, the average speaker is generally very good at estimating the interlocutor's level of involvement, without the need of an explicit verbal feedback. He/she only needs to interpret visually accessible unconscious signals, such as body postures and movements, facial expressions, eye-gazes. The speaker can understand if the addressee is engaged with the discourse, and continuously fine-tune his/her communication strategy in order to keep the communication channel open and the attention high in the audience.[1]

Understanding of non-verbal feedback is not easy to achieve for virtual agents and robots, but this ability is strategic for enabling more natural interfaces capable of adapting to users. Indeed, perceiving signals of loss of attention (and thus, of engagement) is of paramount importance to design naturally behaving virtual agents, enabled to adjust the communication strategy to keep high the interest of their addressees. That information is also a general sign of the quality of the interaction and, more broadly, of the communication experience. At the same time, the ability to generate engaging behaviors in an agent can be beneficial in terms of social awareness (Oertel et al., 2020).

The objective of developing a natural behaving agent, able to guide visitors along a tour in cultural sites, was at the core of the CHROME Project[2] (Cutugno et al., 2018; Origlia et al., 2018), and the present work is intended in the same direction. More specifically, this paper explores the possibility to predict audience engagement in the context of guided tours, by considering acoustic and linguistic features of the speech of an expert guide leading visitors inside museums.

The paper is organised as follows: Section 2 draws a brief overview of related works in the field of engagement annotation and prediction; Section 3 describes in details the construction of the dataset; Section 4 reports the methodology adopted to extract features specific for both linguistic and acoustic modalities; Section 5 illustrates the set of experiments conducted on the collected data, in terms of classification scenarios and features used; Section 6 gathers final observations and ideas for future works.

**Contributions** The main contributions in this paper are: i) a novel multimodal Italian dataset with engagement annotation; ii) multiple classification scenarios experiments; iii) impact of modality-specific features on multimodal classification.



[1] Recent studies have shown that the processing of *emotionality* in prosody, facial expressions and speech content is associated in the listeners' brain with enhanced activation of auditory cortices, fusiform gyri and middle temporal gyri, respectively, confirming that emotional states are processed through modality-specific modulation strategies (Regenbogen et al., 2012).

[2] Cultural Heritage Resources Orienting Multimodal Experience. http://www.chrome.unina.it/

## 2 Related Works

With the word engagement we refer to the level of involvement reached during a social interaction, which assumes the shape of a process through the whole communication exchange. More specifically, Poggi (2007) defines the process of social 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. Another definition, adopted by many studies in Human-Robot Interaction (HRI),[3] describes engagement as the process by which interactors start, maintain, and end their perceived connections to each other during an interaction (Sidner et al., 2005).

Observations and annotations of engagement are collected on the basis of visible cues, such as facial expressions and reactions, eye gazes, body movements and postures. The majority of the studies are often conducted on a dyadic base, i.e. focusing on communication contexts involving only two participants, most of the times a human interacting with an agent/robot (Castellano et al., 2009; Sanghvi et al., 2011; Ben-Youssef et al., 2021). Nevertheless, engagement can be measured in groups of people taking part in the same communication event as the average of the degree to which individuals are involved (Gatica-Perez et al., 2005; Oertel et al., 2011). Human-to-human interactions within groups have been studied principally in the research field of education (Fredricks et al., 2004) where visible cues are related to attention, which is considered as a perceivable proxy to the more complex and inner process of engagement (Goldberg et al., 2019).

## 3 Dataset

The dataset presented in this paper is derived from a subset of the CHROME Project data collection (Origlia et al., 2018), which comprises aligned videos, audios and transcriptions of guided tours in three Charterhouses in Campania. Two videos have been recorded for each session: one video with the guide as subject, the other focused on the group of visitors. Data of 3 visits with the same expert guide (in the same Charterhouse) have been selected. Each visit is organised in 6 points of interest (POI), i.e. rooms or areas inside the Charterhouse where groups stop during the tours and the guide describes the place with its furnishings, history, and anecdotes.

In total, starting data consist of 2:44:25 hours of audiovisual material and 22,621 tokens from the aligned transcriptions. The language of the speech is Italian.

### 3.1 Annotation and Segmentation

Engagement has been annotated as a continuous measurement of visitor's attention, as a visible cue of engagement. The annotation has been carried out using PAGAN Annotation Tool (Melhart et al., 2019), and performed by two annotators watching videos of the groups of visitors in order to observe cues of gain or loss of attention. Following Oertel et al. (2011), annotators have been asked to evaluate the average behaviour of the whole group. Agreement between the two annotators is consistent, with an average Spearman's rho of 0.87 (Ravelli et al., 2020).

The raw transcriptions have been manually segmented with the objective of creating textual segments close to written sentences, and this segmentation has been projected on audio files, in order to obtain aligned text-audio pairs for each segment. Given that every visit is similarly structured, and also topics and whole pieces of information are mostly the same across different visits, the resulting transcriptions are extremely clear and phenomena such as retracting and disfluencies are minimum if compared to transcriptions of typical spontaneous speech. Thus, text normalisation (i.e., disfluencies removal, basic punctuation insertion) has been easy to obtain, and the resulting adaptation lead to sentences easy to parse with common NLP tools trained on written texts.

Segmentation has been performed on the basis of perceptual cues of utterance completeness. As described by Danieli et al. (2005), a break is said terminal if a competent speaker (i.e. mother tongue speaker) assigns to it the quality of concluding the sequence. Starting with this observation, two annotators have been asked to listen to the original audio tracks and mark transcriptions with a full stop where they perceived a break as a boundary between utterances, on the basis of intonation and prosodic contour. Utterances perceived as independent but pronounced too quickly to allow a clean cut (especially considering audio segmentation and the consequent features extraction) have been kept together in a single segment.

[3] For a broad and complete overview of works on engagement in HRI studies, see Oertel et al. (2020)

To assess the reliability of the segmentation process, we measured the accuracy between the two annotators on a subset of the data (the 40% of the total, corresponding to one of the three visits). We adopted a chunking approach to the problem, by adapting an IOB (Inside-Outside-Begin) tagging framework to label tokens, from the continuous transcriptions of the sample, at the beginning (B), inside (I), end (E) of segments, or outside (O) any of those. We measured an accuracy of 91,53% in terms of agreement/disagreement on the basis of the series of labelled tokens derived for each annotator.

At the end of the segmentation process, the dataset counts 1,114 Italian sentences, with an average of 20.31 tokens per sentence (std: 11.96), and an average duration of audio segments of 8.13 seconds (std: 5.22).

An engagement class has been assigned to each sentence: 1 if an increase in engagement has been recorded in the span of that sentence, 0 in case of decrease or no variation. To compute the class, we considered the delta between the input and output values of the continuous measurement obtained with the annotations, with respect to the beginning and end of sentences. Specifically, for each sentence we selected all the annotations (one per millisecond) falling into the sentence boundaries, and then we subtracted the value of the first one from the last one. We reduced the task to a binary classification in order to test to which extent it is possible to predict engaging content before to evaluate the possibility to expand the analysis to a finer classification, accounting also for what is specifically engaging, not-engaging or neutral.

## 4 Features Extraction

In order to train and test a classifier in predicting the engagement of the addressee of an utterance, using both linguistic and acoustic information, features specific for each modality have been extracted independently, and then concatenated as unique vectors representing each entry of the dataset.

### 4.1 Linguistic Features

The textual modality has been encoded by using Profiling–UD (Brunato et al., 2020), a publicly available web–based application[4] inspired to the methodology initially presented in Montemagni (2013), that performs linguistic profiling of a text, or a large collection of texts, for multiple languages. The system, based on an intermediate step of linguistic annotation with UDPipe (Straka et al., 2016), extracts a total of 129 features per each analysed document. In this case, Profiling-UD analysis has been performed per sentence, thus the output has been considered as the linguistic feature set of each segment of the dataset. Table 1 reports the 127 features extracted with Profiling-UD and used as textual modality features for the classifier.[5]

| **Linguistic features** | **n** |
|---:|:---:|
| Raw text properties | 2 |
| Morpho–syntactic information | 52 |
| Verbal predicate structure | 10 |
| Parsed tree structures | 15 |
| Syntactic relations | 38 |
| Subordination phenomena | 10 |
| **Total** | **127** |

Table 1: Set of linguistic features extracted with Profiling-UD.

### 4.2 Acoustic Features

The acoustic modality has been encoded using OpenSmile[6] (Eyben et al., 2010), a complete and open-source toolkit for analysis, processing and classification of audio data, especially targeted at speech and music applications such as automatic speech recognition, speaker identification, emotion recognition, or beat tracking and chord detection. The acoustic features set used in this case is the Computational Paralinguistics ChallengE[7] (ComParE), which comprises 65 Low-Level Descriptors (LLDs), computed per frame. Table 2 reports a summary of the ComParE LLDs extracted with OpenSmile, grouped by type: prosody-related, spectrum-related and quality-related.

Given that the duration (and number of frames, consequently) of audio segments varies, common transformations (min, max, mean, median, std) have been applied on the set of per-frame features

[4]Profiling-UD can be accessed at the following link: `http://linguistic-profiling.italianlp.it`

[5]Out of the 129 Profiling-UD features, *n_sentences* and *tokens_per_sent* (raw text properties) have not been considered, given that the analysis has been performed per sentence.

[6]`https://www.audeering.com/research/opensmile/`

[7]`http://www.compare.openaudio.eu`

| Acoustic features | n |
|---|---|
| ***Prosodic*** | |
| $F_0$ (SHS and viterbi smoothing) | 1 |
| Sum of auditory spectrum (loudness) | 1 |
| Sum of RASTA-style filtered auditory spectrum | 1 |
| RMS energy, zero-crossing rate | 2 |
| ***Spectral*** | |
| RASTA-style auditory spectrum, bands 1–26 (0–8 kHz) | 26 |
| MFCC 1–14 | 14 |
| Spectral energy 250–650 Hz, 1 k–4 kHz | 2 |
| Spectral roll off point 0.25, 0.50, 0.75, 0.90 | 4 |
| Spectral flux, centroid, entropy, slope | 4 |
| Psychoacoustic sharpness, harmonicity | 2 |
| Spectral variance, skewness, kurtosis | 3 |
| ***Sound quality*** | |
| Voicing probability | 1 |
| Log. HNR, Jitter (local, delta), Shimmer (local) | 4 |
| **Total** | **65** |

Table 2: Set of acoustic features extracted with OpenSmile.

of each segment, leading to a total of 325 acoustic features (65 LLDs x 5 transformations).

## 5 Experiments

To explore the possibility to predict engaging sentences, we implemented a machine learning classifier using the linear SVM algorithm provided by the scikit-learn library (Pedregosa et al., 2011).

We defined various classification scenarios on the basis of 3 different train-test splitting of the dataset. The first, and more common scenario, is based on a *k*-fold setting, in which data has been randomly split in 10 folds, trained on 9 of them and tested on the remaining one. The second scenario uses data from one POI from all the visits as a test, and it is trained on the remaining parts. The third scenario considers data from a whole visit as test and is trained on the remaining two. Global results are obtained by averaging the classification performances of each run per scenario (e.g. average of all *k*-fold outputs tested on every fold).

For each scenario, the SVM classifier has been trained and tested three times, once per single modality (i.e. linguistic or acoustic features exclusively) and once with joint representations (the full set of both linguistic and acoustic features). All the features have been normalised in the range $[0, 1]$ using the *MinMaxScaler* algorithm implemented in scikit-learn.

| | ***k*-fold** | **POI** | **Visit** |
|---|---|---|---|
| Baseline | 51.53% | 47.05% | 47.32% |
| Linguistic | **57.81%** | **58.05%** | **57.44%** |
| Acoustic | 55.35% | 55.64% | 55.83% |
| Multimodal | 53.49% | 54.25% | 54.40% |

Table 3: Accuracy scores for each classification scenario with all features settings.

Table 3 reports the aggregated results, in terms of accuracy, from all the experiments. The baseline considered is the assignment of the majority class found in the training data. All the classifiers in the three scenarios obtain better results than the baseline, but the multimodal systems (the ones exploiting both linguistic and acoustic sets of features) are never able to do better than models based on linguistic features only. Moreover, it is possible to observe that multimodal systems achieve scores similar to acoustic systems.

Low performances, especially for multimodal systems, may be ascribed to the fact that the classifiers are fed with too many features (452 total; 127 textual and 325 acoustic features) with respect to the dimension of the dataset (1,114 items), and thus they build representations with low variation in terms of single feature weight. Moreover, summing the two sets in the multimodal systems leads to worst results than single-modality systems, amplifying the problem.

In order to verify this hypotesis, we reduced the number of features by observing the weights assigned to each feature by classifiers trained on single modalities, and selecting only the top 20 from each ranked set. Figures 1 and 2 show the reduced set of features along with their weights for the linguistic and acoustic set of features, respectively. Among the top-rated, on the linguistic side, we can find features related to the syntactic tree of the sentence and verbal predicate structure; on the acoustic side, principally spectral and prosodic features.

As shown in Table 4, by using this reduced features sets, all systems obtain better results with respect to the experiments conducted exploiting the whole sets of features. Most significant improvements can be traced for models based on acoustic

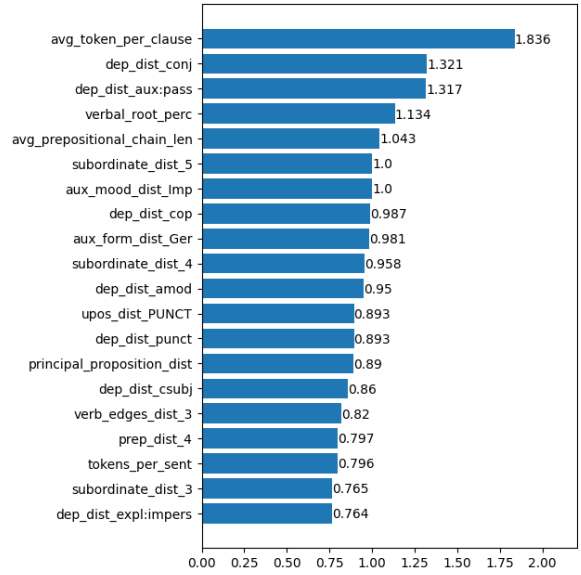


Figure 1: Top 20 linguistic features.

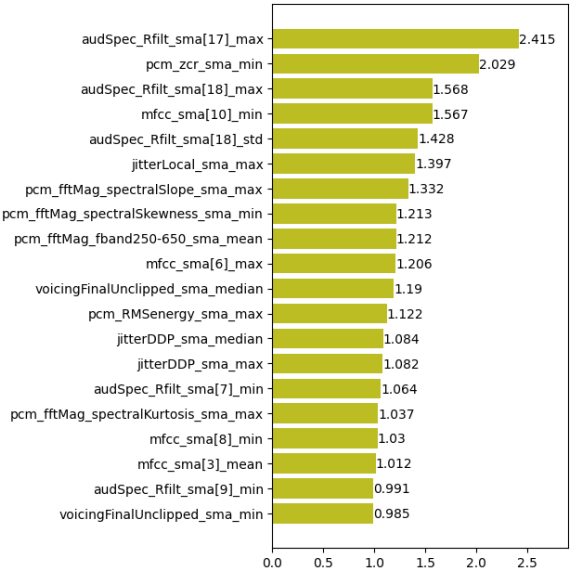


Figure 2: Top 20 acoustic features.

| | ***k*-fold** | **POI** | **Visit** |
|---|---|---|---|
| Baseline | 51.53% | 47.05% | 47.32% |
| Linguistic | 60.78% | 59.86% | 60.39% |
| Acoustic | 65.70% | 63.87% | **64.86%** |
| Multimodal | **66.07%** | **65.36%** | 64.03% |

Table 4: Accuracy score for each classification scenario with best features settings.

and multimodal features set, with an average increase in accuracy of the 10%. Differently from previous experiments, multimodal systems reach the best overall results in two out of three scenarios (*k*-fold and POI).

Again, multimodal systems scores are close to those obtained exploiting exclusively acoustic features. For this reason, we compared the predictions from single modalities with multimodal ones, and we found out that multimodal systems predictions overlap more with acoustic systems (0.86) than with linguistic systems (0.79). It confirms that this behaviour is due to the fact that acoustic features are those more considered by the multimodal classifier.

It is possible to observe the higher contribution from acoustic features to the multimodal systems in Figure 3: among the top 10 most important features, only 2 are linguistic, and the trend is dramatically off balance in favour of acoustic features.

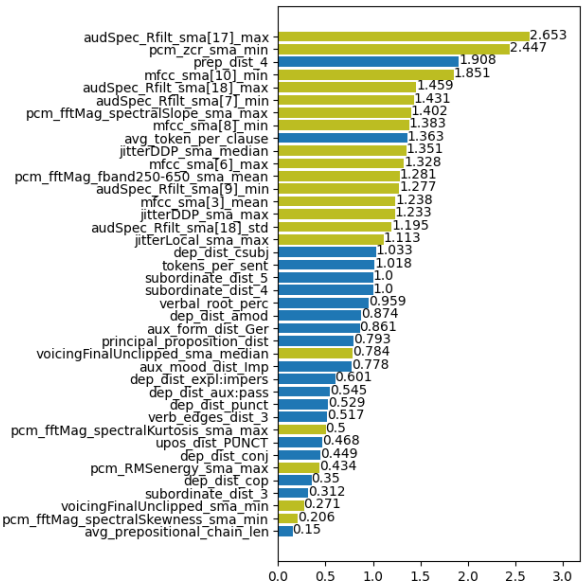


Figure 3: Features weight after selecting best 20 linguistic and acoustic features.

## 6 Conclusions

In this paper we introduced a novel multimodal dataset for the analysis and prediction of engagement, composed of Italian sentences derived from the speech of an expert guide leading visitors in cultural sites, enriched with multimodal features, and labelled on the basis of the perceivable engagement of the audience. We performed several experiments in different classification scenarios, in order to explore the possibility to predict engage-

ment on the basis of features extracted for both the linguistic and acoustic modalities. Combining modalities in classification leads to good results, but with a filtered set of features to avoid too noisy representations. An in interesting experiment would be to combine the outcomes of two different systems (one exploiting exclusively acoustic features, linguistic features the other) rather than using a monolithic one fed with all the features. This technique often leads to better performances with respect to the decisions taken by a single system (Woźniak et al., 2014; Malmasi and Dras, 2018).

Moreover, we are working on aligning features derived from the visual modality, by encoding information from the videos used to annotate engagement. In this way, the dataset will contain a more complete representation, and it would be possible to correlate perceived engagement in the audience with the full set of stimuli offered during the guided tour.

## Acknowledgments

The authors would like to acknowledge the contribution of Luca Poggianti and Mario Gomis, who have annotated the engagement on the videos, and Federico Boggia and Ludovica Binetti, who have segmented the sentences of the dataset.

## References

Atef Ben-Youssef, Chloé Clavel, and Slim Essid. 2021. Early detection of user engagement breakdown in spontaneous human-humanoid interaction. *IEEE Transactions on Affective Computing*, 12(3):776–787.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.

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. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 119–126.

Francesco Cutugno, Felice Dell'Orletta, Isabella Poggi, Renata Savy, and Antonio Sorgente. 2018. The CHROME Manifesto: Integrating Multimodal Data into Cultural Heritage Resources. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics, CLiC-it 2018, Torino, Italy, December 10-12 , 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Morena Danieli, Juan María Garrido, Massimo Moneglia, Andrea Panizza, Silvia Quazza, and Marc Swerts. 2005. Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM". In Emanuela Cresti and Massimo Moneglia, editors, *C-ORAL-ROM: integrated reference corpora for spoken romance languages*, pages 1513–1516.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.

Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest-level in meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–489. IEEE.

Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2019. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review*, 35(1):463–23.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Comput. Linguistics*, 44(3).

David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video Affect Annotation Made Easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE.

Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, XLII(1):145–172.

Catharine Oertel, Stefan Scherer, and Nick Campbell. 2011. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Twelfth annual conference of the international speech communication association*.

Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, 7:92.

Antonio Origlia, Renata Savy, Isabella Poggi, Francesco Cutugno, Iolanda Alfano, Francesca D'Errico, Laura Vincze, and Violetta Cataldo. 2018. An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the chrome project. In *2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*, volume 2091.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Isabella Poggi. 2007. *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.

Andrea Amelio Ravelli, Antonio Origlia, and Felice Dell'Orletta. 2020. Exploring Attention in a Multimodal Corpus of Guided Tours. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Christina Regenbogen, Daniel A Schneider, Raquel E Gur, Frank Schneider, Ute Habel, and Thilo Kellermann. 2012. Multimodal human communication — Targeting facial expressions, speech content and prosody. *NeuroImage*, 60(4):2346–2356.

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 *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, page 305–312, New York, NY, USA. Association for Computing Machinery.

Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Michał Woźniak, Manuel Graña, and Emilio Corchado. 2014. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17. Special Issue on Information Fusion in Hybrid Intelligent Fusion Systems.