Attachment Recognition in School Age Children Based on Automatic Analysis of Facial Expressions and Nonverbal Vocal Behaviour

Huda Alsofyani
Huda.Alsofyani@glasgow.ac.uk
University of Glasgow
Glasgow, UK

Alessandro Vinciarelli
Alessandro.Vinciarelli@glasgow.ac.uk
University of Glasgow
Glasgow, UK

ABSTRACT

Attachment is a psychological construct that accounts for whether children are secure (the parents meet their physical and emotional needs) or insecure (the parents do not meet their physical and emotional needs). Unless identified and supported early enough, insecure children develop higher chances of experiencing issues such as antisocial behaviour or suicidal tendencies. For this reason, this article proposes a multimodal approach for attachment recognition in school age children (5-9 years old). In particular, the approach infers the attachment condition of a child from facial expressions and nonverbal vocal behaviour. The experiments involved 104 children that were recorded while undergoing the Manchester Child Attachment Story Test, an instrument that child psychiatrists use often to identify insecure children. The results show that attachment can be recognized with accuracy up to 71.2% (F1 score 62.4%).

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing → Health informatics**;

KEYWORDS

Attachment, Social Signal Processing, School Age Children

ACM Reference Format:
Huda Alsofyani and Alessandro Vinciarelli. 2018. Attachment Recognition in School Age Children Based on Automatic Analysis of Facial Expressions and Nonverbal Vocal Behaviour. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

According to child psychiatrists, the way parents (or more generally caregivers) address the needs of children leaves traces in terms of *attachment*, a psychological construct that accounts for whether "*the infant's search for consistent care is met with either success, leading to a sense of emotional security, or failure, with insecurity as a result*" [23]. In particular, children are said to be *secure* or *insecure* depending on whether they perceive caregivers to be responsive or not, respectively [43]. While resulting from early childhood interactions, the attachment condition shapes the perception of relationships during the whole life of an individual [7]. As a consequence, once they become adult, insecure children tend to have more problems with colleagues, friends and romantic partners [23]. This leads to lower levels of satisfaction in social, family and professional life, respectively. Furthermore, the lack of care and response that insecure individuals experience in childhood reinforces stress responses that, in adult life, increase the chances of, e.g., coronary pathologies [9, 29] or antisocial tendencies [44].

The best way to address the issues above is to identify and support insecure children as early as possible. For this reason, this article proposes a multimodal approach aimed at inferring the attachment condition of school-age children (5 to 9 years old) from facial expressions and nonverbal vocal behaviour. The approach processes independently these two modalities and then applies a fusion scheme to combine the decisions made separately for each of them. Experiments and results show that the multimodal combination improves over both unimodal approaches to a statistically significant extent.

The experiments involved 104 children randomly recruited among primary school pupils. Every child was recorded while undergoing the *Manchester Child Attachment Story Task* (MCAST) [15], a test that child psychiatrists commonly apply to assess the attachment condition of children. During the MCAST, the participants describe everyday life interactions between children and their mothers (see Section 3 for more details). Assessors identify the attachment condition through the way children perform such a task. Therefore, it is common clinical practice to record the administration of the MCAST in order to analyze the behaviour of children in detail. Such an approach naturally lends itself to the application of Social Signal Processing (SSP), the computing domain aimed at inferring social and psychological phenomena from nonverbal behaviour [41]. In particular, the experiments show that SSP methodologies lead to an accuracy of 64.6% (F1 Score 56.2%) for facial expressions, 68.9% (F1 Score 59.6%) for nonverbal vocal behaviour and 71.2% (F1 Score 62.4%) for their combination.

To the best of our knowledge, this is one of the earliest works aimed at the recognition of attachment in school-age children (see Section 2). As a consequence, the literature does not provide indications on behavioural modalities most likely to convey attachment-relevant information. The proposed approach is based on facial expressions and nonverbal vocal behaviour because these modalities were shown to effectively account for a wide range of social and psychological phenomena (see, e.g., [36, 41]). The results of

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

the work appear to confirm such an observation because both unimodal and multimodal approaches perform well above chance. This suggests that children, while undergoing the MCAST, actually manifest their attachment condition through facial expressions and nonverbal vocal behaviour. In addition, the results show that the multimodal approach improves over unimodal ones to a statistically significant extent. This suggests that face and speech tend to convey complementary information, a condition necessary for two modalities to mutually improve each other [33].

The rest of this article is organized as follows: Section 2 provides a survey of previous work, Section 3 describes attachment assessment and data collection, Section 4 presents the proposed approach, Section 5 reports on experiments and results and the final Section 6 draws some conclusions.

2 PREVIOUS WORK

Mental health issues have attracted significant attention in the computing community, especially in regard to "*systems designed for use in prevention of mental illness, standalone computer-based treatment and self-help systems, and systems intended for use in conjunction with face-to-face psychotherapy [...] monitoring of and self-monitoring by clients, communication (such as computed mediated therapy), delivery of content [...] and interaction with content [...]*" [8]. In the case of pathologies such as depression or autism, the literature proposes a large number of approaches aimed at automatic detection and diagnosis, while in the case of attachment, most efforts focus on different problems.

In several cases, attachment-relevant works aim at ensuring positive relationships between users and technology, in particular social robots [17–19] and interactive systems [24, 25, 27, 34, 42]. The core-assumption underlying such works is that attachment plays a role not only in the interaction between people, but also in the interaction between people and machines. This is particularly evident in the case of social robots expected to interact with people like people do with one another [5]. The experiments in [17] show that the more a robot is sophisticated, the more it is effective at establishing attachment bonds with its users. In a similar vein, the experiments in [18, 19] show that robots simulating attachment-related behaviours can help their users to develop parenting styles likely to foster secure attachment. Similar effects were observed for digital artefacts not designed to replicate the way people behave [24, 25]. The result was a design theory aimed at ensuring that users, in particular children with limited attention span, establish longer term relationships with the technologies they use [27, 34]. As an indirect confirmation, the experiments in [42] have shown in quantitative terms that secure children tend to respond to a software system in a way that is more coherent with the way the system is designed.

Other attachment-related works support the development of secure attachment through the use of technologies for communication between parents and children [13, 16, 20], based on the assumption that it is quality of interaction that shapes the attachment condition of children. The work proposed in [13] suggests that relationships between parents and children can be enhanced, from an attachment point of view, through the use of stickers capable to emit simple stimuli. Parents and children can design the stimuli and decide to what objects the stickers must be applied. The focus in [16] is on helping parents to better communicate with deaf children, especially in regard to telling stories. The intention of the authors is to increase the chances of secure attachment despite the hearing difficulties. Finally, the experiments in [20] aim at showing that mobile technologies can provide a good communication environment for families. In this way, children have higher chances of developing secure attachment.

Only a few works presented in the literature addressed the problem of recognizing attachment based on observable evidence. The earliest works showed that there is a relationship between attachment and blood pressure measured through ear pulse waves [28, 40]. While not being an attempt to perform attachment recognition, these results still provided an initial indication that attachment leaves machine detectable traces. Actual attachment recognition was addressed only recently for adults and children. In the first case [30], the proposed approach was based on photoplethysmography, facial behaviour, paralinguistics and language. The best result was a Root Mean Square Error of 12.1 in predicting the scores obtained with an attachment self-assessment questionnaire. In the second case [35], deep networks were fed with a representation of the way children move and the best result was an accuracy of roughly 75% in discriminating between secure and insecure children.

Overall, this state-of-the-art suggests that the computing community has addressed the problem of attachment only to a limited extent. In most cases, the focus was on improving the interaction between users and technology or the interaction between people through technology. Only a few works, to the best of our knowledge, tried to infer the attachment condition of an individual from observable evidence (verbal or nonverbal behaviour and language). One possible reason is that the computing community has focused on issues such as depression and autism because these impact the wellbeing of an individual in a more evident way. However, as the long-term effects of insecure attachment become increasingly more evident [23], attachment recognition might attract more attention in the next years.

3 DATA COLLECTION

The *Manchester Child Attachment Story Task* (MCAST) [15] is one of the tests that child psychiatrists use most frequently to assess the attachment condition of children. During the MCAST, participants listen to five short *story stems* about a child and her mother:

- *Breakfast* (the child wakes up in the morning and the mother prepares breakfast);
- *Nightmare* (the child wakes up after a nightmare and calls the mother for comfort);
- *Hopscotch* (the child gets a wound on her knee and asks the mother to provide first aid);
- *Tummyache* (the child feels a pain in the stomach and asks the mother to provide assistance);
- *Shopping* (the child looses contact with the mother in a shopping mall and tries to re-establish contact with her).

At the end of each stem, the participants have to explain how the story continues with the help of two dolls corresponding to the two main characters (*baby doll* and *mummy doll*). The key-assumption

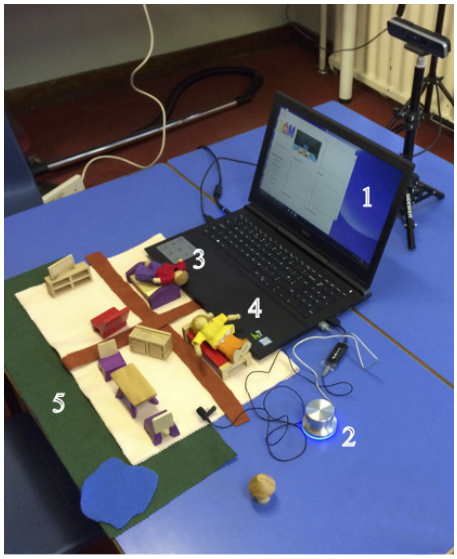

Figure 1: The picture shows the MCAST administration system used in the experiments. Element 1 corresponds to a laptop screen where the participants can watch videos where an actor delivers the story stems and provides guidance about the steps of the test. Element 2 is a button that the participants are asked to push every time they complete an MCAST step. Elements 3 and 4 are the dolls that participants use to represent the continuation of the story stems. Element 5 is the toy house that helps the children with the story stem representation.

is that narrative tasks are "*[...] a vehicle for accessing the representational world of young children through the developmentally appropriate domain of play, and these techniques have been used successfully in investigations of representation and attachment [...]*" [39]. In other words, while performing an activity that looks like a game because of the dolls, children manifest their attachment condition through the way they tell the continuation of the stems. The expectation is that children in different attachment condition tend to represent the stories in a different way.

Figure 1 shows the *School Attachment Monitor* (SAM), the apparatus used to administer the test and record the children. The administration protocol was based on the following main steps:

- *Story stem delivery*: the SAM plays a video on a computer screen (element 1in Figure 1) where an actor delivers a story stem and, at the end, prompts the participants to continue the story by using the dolls (elements 3 and 4 in Figure 1) and the play mat representing an apartment (element 5 in Figure 1);
- *Story representation*: the participants, while being recorded with a camera and a microphone, continue the story stem and, at the end, they press the "*Finish*" button (element 2 in Figure 1);
- *Iteration*: if the story stem is one of the first four, the system goes back to the first step, otherwise it concludes the test.

	P1 (5-6)	P2 (6-7)	P3 (7-8)	P4 (8-9)
F	9 (8.6%)	22 (21.1%)	15 (14.5%)	11 (10.6%)
M	10 (9.6%)	18 (17.3%)	14 (13.5%)	5 (4.8%)
S	9 (8.6%)	22 (21.1%)	18 (17.3%)	10 (9.6%)
I	10 (9.6%)	18 (17.3%)	11 (10.6%)	6 (5.9%)
Tot.	19 (18.2%)	40 (38.4%)	29 (27.9%)	16 (15.5%)

Table 1: The table shows how the 104 participants distribute across different school levels, from Primary 1 (P1) to Primary 4 (P4). For every level, the table provides gender distribution (*F* and *M* stand for female and male, respectively) and attachment condition distribution (*S* and *I* stand for secure and insecure, respectively).

A pool of four assessors that attended the training course to become professional MCAST assessors [14] examined the videos collected after every administration of the test. In particular, the assessment was performed according to common clinical practice: two random members of the pool assess independently the same child and, if there is agreement, the assessment is accepted, otherwise, all members of the pool discuss and achieve a consensual decision.

In total, the experiments involved 104 children randomly recruited in different primary schools. The total lenght of the recordings is 18 hours, 30 minutes and 34 seconds (the average is 640.7 seconds per child) and Figure 2 shows the amount of material for every individual participant. The average length of the videos for the different story stems is 137.0 ± 78.3 for *Breakfast*, 117.6 ± 73.6 for *Nightmare*, 116.3 ± 66.4 for *Hopscotch*, 111.2 ± 58.2 for *Tummyache* and 158.4 ± 78.5 for *Shopping Mall*. The difference between this latter stem and the others is statistically significant ($p < 0.05$ according to a two-tailed t-test with unequal variance). Similarly, the *Breakfast* narratives are longer, to a statistically significant extent, than those of the stems with a smaller average length. In the case of the remaining three narratives, there is no statistically significant difference between the respective average durations.

One possible reason for *Breakfast* leading to longer narratives is that, being the first story of the MCAST, it is used to provide information that does not need to be repeated for the other stems (e.g., an overall assessment of the relationship between the story child and the mother). For what concerns *Shopping Mall*, the probable reason for leading to the longest narratives, on average, is probably that the stem does not take place in the same setting as the others (the apartment corresponding to the play mat in Figure 1). Therefore, children are likely to need extra-time to introduce information about the differences with respect to the other stems.

Table 1 provides demographic information about the 104 children involved in the experiments. In terms of gender distribution, there are 57 female participants and 47 male participants, corresponding to 54.8% and 45.2% of the total, respectively. For what concerns attachment, there are 59 secure participants and 45 insecure participants, corresponding to 56.5% and 43.4% of the total, respectively. According to a χ^2 test, there is no statistically significant difference with respect to the attachment distribution observed in the general population [10, 26]. The data were collected according to the ethical regulations of the country where the experiments were conducted.

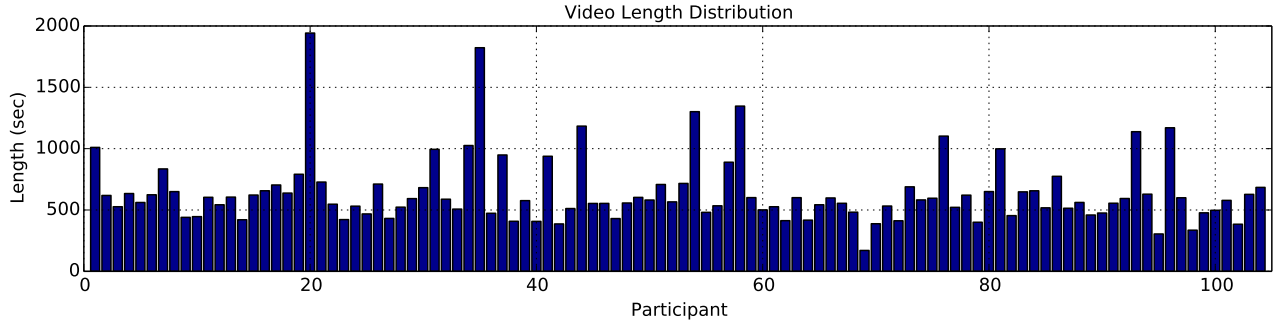

Figure 2: The chart shows, for every participant, the total lenght of the videos at disposition.

In particular, children were involved in the experiments only upon written authorization of their parents. Furthermore, they were allowed to interrupt the participation whenever they did not feel comfortable. Parents were given access to the data of their children and were given the possibility to destroy them partially or totally.

4 THE APPROACH

This section presents the two unimodal approaches used in the experiments and the way their outcomes were combined.

4.1 Face-Based Unimodal Recognition

Section 3 shows that the School Attachment Monitor, the apparatus used for collecting the data, records children undergoing the MCAST. The videos make it possible to analyze the facial expressions of children and the proposed face-based approach includes three main steps, namely *feature extraction*, *recognition* and *aggregation*. The feature extraction step starts by extracting a vector $\vec{f}$ of dimension $D = 17$ from every video frame. The components f_i correspond to the activation intensity of D Action Units (AU) extracted with *OpenFace* [2], a publicly available package for facial behaviour analysis:

- *Eyes Area*: Inner brow raiser (AU1), outer brow raiser (AU2), brow lowerer (AU4), upper lid raiser (AU5), cheek raiser (AU6), lid tightener (AU7), blink (AU45);
- *Nose Area*: Nose wrinkler (AU9);
- *Mouth Area*: Upper lip raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), lip stretched (AU20), lip tightener (AU23), lips part (AU25), jaw drop (AU26).

Given a training set, it is possible to identify, for each of the AUs above, the top 5% intensity values that were observed. Such values account, at least for a particular AU, to the biggest differences with respect to a neutral expression. Correspondingly, for every AU, it is possible to identify a threshold θ that discriminates between AU intensity values in the top 5% (meaning that they are above the threshold) and the others.

Thanks to the θ thresholds above, it is possible to calculate the fraction of frames in a video such that the intensity of a particular AU is in the top 5% of the training set. This leads, for every video, to a feature vector where the component i is the percentage of frames in which the intensity of the i^{th} AU is above the corresponding θ threshold. The feature vectors obtained through such a process allow one to train a classifier that can discriminate between secure and insecure children, a task that corresponds to the recognition step of the approach. In the experiments of this work the classifier is a Logistic Regression.

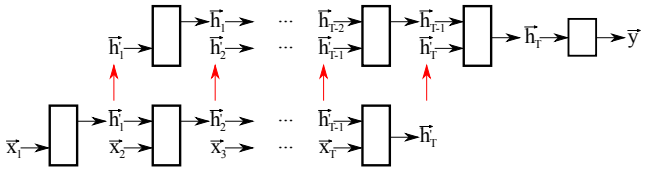

Figure 3: The figure shows how information flows through the stacked RNNs. Vectors $\vec{h}_t$ and $\vec{h}'_t$ correspond the hidden states of the two RNNs, while the $\vec{x}_t$s are the feature vectors extracted from the speech data. Rectangular blocks represent hyperbolic tangent layers, while the square block is a softmax layer.

Once the classifier has been trained, it is possible to classify a child as secure or insecure and, given that there are 5 different videos per child (one per MCAST story stem), there are 5 classification outcomes. In particular, the Logistic Regression provides, for every video, the two posterior probabilities corresponding to classes *secure* and *insecure*. This allows one to assign a child to the class corresponding to the highest average posterior probability, an aggregation approach referred to as *Weighted Average* (WA).

4.2 Speech-Based Unimodal Recognition

The unimodal speech-based approach includes three main steps, namely *feature extraction*, *recognition* and *aggregation*. The first step relies on *OpenSmile* [11, 12], a publicly available software package for speech processing. The signal is first segmented into 33 ms long non-overlapping windows (every window corresponds to a frame in the video) and then, the content of every window is converted into a vector of dimension $D = 32$ that includes 16 features and the respective delta coefficients (these latter are the differences between the value of the feature in the current window and the value of the feature in the previous window). The 16 features are as follows:

- *Root mean square of the energy* (1 feature): it accounts for how loud children speak;

- *Mel Frequency Cepstral Coefficients* (12 features): it accounts for the phonetic content of speech;
- *Zero Crossing Rate* and *Fundamental frequency* (2 features): they account for thefrequency on which most of the ignal energy concentrates and contributes to define the way a voice sounds;
- *Voicing probability* (1 feature): it accounts for the presence of silences.

The motivation behind the choice of the features above is that they were designed for an emotion recognition challenge [37] and, since then, they were shown to be effective for the inference of a wide spectrum of social an psychological phenomena from speech.

The feature extraction process converts the speech signals into sequences $X = (\vec{x}_1, \ldots, \vec{x}_T)$ of feature vectors that are given as input to the recognition step. In particular, the sequences X are fed to two *stacked* Recurrent Neural Networks (RNNs) [31], i.e., two RNNs connected in such a way that the hidden states of the first one act as input to the second one (see Figure 3). The motivation behind the choice of RNNs is that such models are suitable for processing sequential information (the sequences X account for changes over time of speech signal properties). Furthermore, the use of a stacked architecture is expected to capture higher level of abstraction, i.e., how the changes captured through the first RNN change over time.

One of the main problems in training stacked RNNs is that gradients can vanish or explode when the input sequences are too long [32]. For this reason, the input sequences X were split into non-overlapping segments including to $L = 128$ vectors each (corresponding to roughly 4.25 seconds of speech). This is a tradeoff between the need to have sequences long enough to capture long-term temporal information, but short enough to avoid training issues. The model was trained to provide as output the probability of one of the segments above being uttered by an insecure child (through the softmax layer the second RNN is connected to). Given that the sequences X include multiple segments of 128 vectors, there are multiple decisions for every speech recording. For this reason, a speech recording is assigned to the class its segments are most frequently assigned to. In other words, if f_s and f_i are the fractions of segments assigned to class secure and insecure ($f_s + f_i = 1$), the recording is assigned to class $\hat{c} = \arg\max_{c=i,s} f_c$.

Like in the case of the unimodal face-based approach, the recordings corresponding to the MCAST stems are classified individually and, therefore, there are five classification outcomes per child. This requires an aggregation step that is performed like in the case of the face-based approach (see end of Section 4.1), namely through Weighted Average. In particular, children are assigned to the class $\hat{c}$ corresponding to the highest average value of $f_{\hat{c}}$, where f_c is the fraction of segments assigned to class c in a recording (see above).

4.3 Multimodal Combination

The unimodal approaches presented in this section classify every recording as being produced by a secure or insecure child, respectively. Given that, for every child, there are five recordings and two modalities, this means that there are 10 classification outcomes per child. The combination is performed using the same approaches applied for the unimodal recognizers, i.e., the Weighted Average (WA). The scores output by the various classifiers for all recordings assigned to a class c are averaged. The child is then assigned to the class for which the average is greater.

5 EXPERIMENTS AND RESULTS

The experiments were performed according to a k-fold protocol ($k = 10$). The 104 participants were split into k disjoint groups through a random process and every fold was created by using all the data corresponding to the participants of a group. In such a way, the experiments were *person independent*, i.e., the participants were never represented in both training and test set. This ensures that the approach actually recognizes the attachment condition of the children and not their identity. Given that the training process involves a random component (initialization of the RNNs and partitioning of the data into folds), every experiment was repeated $R = 10$ times. For this reason, all results are presented in terms of average and standard deviation of different performance metrics over the R repetitions.

In the case of the unimodal face-based appoach (see Section 4.1), the recognition is performed using a Logistic Regression [3]. Such a model does not require one to set any hyperparameters and, therefore, there was no need to perform crossvalidation. In the case, of the speech-based unimodal approach, there are multiple hyperparameters to be set, but they were all given values considered to be standard in the literature. As a consequence, no crossvalidation was performed either. In particular, the dimension of the hidden states was set to $D = 70$, the learning rate to 10^{-3} and the number of training epochs to $T = 50$. The training was performed according to a *mini-batch* strategy aimed at limiting computational issues [22]. Correspondingly, the RNNs were trained over subsets of the training material including $B = 512$ sequences each (the union of all mini-batches corresponds to the whole training set and all mini-batches are disjoint). The risk of overfitting was limited by applying L2 regularization (the λ parameter was set to 10^{-2}).

Table 2 shows the attachment recognition results for the individual story stems (the results were obtained by training and testing over the material corresponding to one of the stems) and for their combination (the results were obtained through WA by aggregating the results obtained for the individual stems). The best Accuracy and F1 score (71.2% and 62.4%, respectively) were obtained with the multimodal combination based on WA. The improvement with respect to the best unimodal results is statistically significant (according to a two-tailed t-test with unequal variance). The baseline for comparison is a random system that assigns a child to class c with probability $p(c)$ corresponding to its prior (see lowest line of Table 2). All approaches improve over such a baseline to a statistically significant extent. This confirms that attachment leaves traces sufficiently consistent to allow automatic detection in both modalities, at least for the 104 children involved in the experiments.

The effectiveness of the unimodal approaches changes significantly from one story stem to the other. In the case of the face-based recognizer, the best performances have been obtained for *Nightmare* and *Tummyache* (the accuracies for these two stems are higher, to a statistically significant extent, than those observed for the others). In the case of the speech-based approach, the best results were obtained for *Breakfast*, *Shopping Mall* and *Hopscotch* (the difference with respect to the other stories is statistically significant). One

Story	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Face-Based				
Breakfast	59.4±1.2	52.9±1.4	51.9±2.5	52.3±1.6
Nightmare	61.9±2.3	56.5±2.9	52.2±3.2	54.3±3.0
Tummyache	61.4±2.8	55.9±2.9	49.8±2.5	52.6±2.4
Hopscotch	57.4±2.0	50.2±2.7	44.3±3.1	47.0±2.3
Shop. Mall	58.9±2.1	51.9±2.5	50.5±4.4	51.1±3.4
All (WA)	64.6±1.2	60.6±2.0	52.4±2.1	56.2±1.2
Speech-Based				
Breakfast	65.8±2.7	63.9±4.1	47.3±8.0	54.0±5.9
Nightmare	61.0±3.6	56.7±5.9	41.8±6.4	47.9±5.6
Tummyache	60.1±4.5	54.5±6.3	46.9±6.8	50.3±6.0
Hopscotch	64.2±4.4	60.3±6.4	47.3±8.2	52.7±7.1
Shop. Mall	65.3±2.6	62.6±5.1	48.5±5.0	54.4±3.3
All (WA)	68.9±2.0	67.8±2.4	53.3±5.0	59.6±3.8
Multimodal				
Breakfast	66.8±2.5	62.7±3.8	56.0±3.4	59.1±2.5
Nightmare	66.0±3.4	62.5±4.0	52.7±7.3	57.0±6.0
Tummyache	64.8±2.3	61.1±3.0	50.7±4.7	55.3±3.7
Hopscotch	63.4±2.0	59.9±3.8	44.1±3.4	50.7±2.7
Shop. Mall	62.0±2.5	57.5±3.6	42.0±5.1	48.5±4.5
All (WA)	71.2±1.6	71.5±2.4	55.6±4.9	62.4±3.1
Random Baseline				
Random	51.0	43.0	43.0	43.0

Table 2: This table shows the performance of the proposed approach in terms of Accuracy (Acc.), Precision (Pre.), Recall (Rec.) and F1 score (F1). The performance metrics are reported in terms of average and standard deviation over 10 repetitions (at every repetition, the RNNs have been initialized differently and the data were split differently for the k-fold). The acronym WA stand for Weighted Average. The Random classifier assigns samples to classes according to a-priori probabilities.

probable explanation is that children react differently to different stems and, therefore, some of these elicit detectable attachment behaviours more frequently than others among the experiment participants.

One interesting aspect of the results in Table 2 is that the unimodal approaches tend to achieve their best performances over different story stems (see above). In particular, the highest peformances of one unimodal approach correspond to the lowest performances of the other one and viceversa. Such a complementarity probably accounts for the tendency to make different mistakes over different samples, a property referred to as *diversity* [33] and known to increase the chances of a classifier *ensemble* [21] to perform better than its best individual member. This is the probable reason why, for the individual story stems, the multimodal combination performs at least as well as the best unimodal approach in 4 cases out of 5 (the only exception is *Shopping Mall*). Furthermore, the diversity is likely to explain why the aggregation of the decisions

Level	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
P1	64.7±4.3	71.8±4.3	54.0±7.0	61.5±6.0
P2	67.8±2.8	70.5±4.4	48.9±6.3	57.5±4.9
P3	75.9±2.3	72.9±4.8	59.1±9.8	64.6±6.1
P4	78.8±5.3	73.4±11.7	71.7±13.7	71.4±8.0

Table 3: The table shows the performance at level Primary 1 (P1) to Primary 4 (P4). See Table 2 for the metrics.

made by the multimodal approach for the individual stems leads to the best results.

Section 3 shows that the 104 experiment participants distribute over primary school levels corresponding to different ages, from P1 (age-range 5 to 6) to P4 (age range 8-9). This means that the children are likely to be at different development stages [6] and, correspondingly, they display different levels of compliance and proficiency in undergoing the MCAST. For this reason, Table 3 shows how the performance of the WA multimodal approach changes for children belonging to different school levels (the analysis was done for such approach because it has the best performance). The difference between P3 and P4 is not statistically significant and, therefore, the approach appears to perform with the same effectiveness for the children of these two levels. However, the difference between P3 and the others is statistically significant, thus suggesting that the performance of the approach tends to improve for older children. In particular, when taking into account only the children at levels P3 and P4, the Accuracy is 76.9%, the Precision is 73.1%, the Recall is 63.6% and the F1 Score is 67.0%.

Another factor that can interplay with the effectiveness of the approach is the amount of data available for individual children. Section 3 shows that the length of the recordings changes significantly from one participant to the other (see Figure 2). Therefore, it is possible to expect that the chances of correctly identifying the attachment condition of one child depend on how long her or his recordings are. For this reason, the children have been split into two groups, namely those that the multimodal approach has correctly classified in all 10 repetitions of the experiments (54 out of 104) and those that the same approach has recognized correctly only in some of the repetitions (50 out of 104). The average length of the recordings is 631.8 ± 236.7 *sec* in the first case and 650.6 ± 312.02 *sec* in the second case. According to a two-tailed t-test with unequal variance, such a difference is not statistically significant and this suggests that there is no association between the length of the material available for a participant and the chances that this latter is classified correctly.

One possible explanation of the result above is that the behavioural traces of attachment are sufficiently consistent to be detected irrespectively of the amount of data at disposition, at least to a certain extent. However, an alternative explanation is that there is enough material for each child to allow automatic attachment recognition. In particular, Figure 2 shows that the shortest recording length for a child is 170.2 seconds (close to 3 minutes), a length that might be above the threshold (if any) necessary for attachment assessment. In other words, it is not possible to exclude that the

Gender	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Face-Based				
Female	71.4±2.5	74.9±3.9	56.2±2.7	64.2±3.0
Male	56.4±1.8	46.2±2.1	47.4±4.3	46.7±2.3
Speech-Based				
Female	66.5±4.2	66.0±5.4	55.0±7.9	59.8±6.0
Male	71.7±3.5	71.7±6.7	50.0±5.7	58.8±5.1
Multimodal				
Female	72.5±2.6	74.4±3.2	60.4±5.1	66.6±3.8
Male	69.6±3.2	66.9±4.8	48.9±8.6	56.2±6.6

Table 4: This table shows the performance of all approaches (unimodal and multimodal) for female and male participants separately.

corpus used in the experiments does not include children that do not speak enough to provide sufficient material.

The last factor that can possibly interplay with the effectiveness of the approach is gender. For this reason, table 4 shows the performance of both unimodal and multimodal approaches for female and male participants separately. In all cases, the performance difference is statistically significant ($p < 0.05$ according to a two-tailed t-test with unequal variance), thus suggesting that there is an association between gender and recognition effectiveness. Given that the approach tends to be more effective for later school levels (see above), one possible explanation is that experiment participants of different gender distribute differently across school levels (see Table 3). However a χ^2 test shows that this is not the case and, therefore, such an explanation is not valid. In a similar vein, it is possible that the differences of Table 4 depend on a different gender distribution across attachment conditions. However, a χ^2 shows, once again, that this is not the case. Therefore, the results of Table 4 are likely to depend on actual differences in the way of expressing attachment between female and male participants. In particular, female children seem to express their condition more reliably through facial expressions, while male ones seem to do it through nonverbal vocal behaviour.

6 CONCLUSIONS

This article presents experiments on automatic attachment recognition in school-age children. The proposed multimodal approach takes into account two behavioural streams (facial expressions and nonverbal vocal behaviour) and automatically recognizes whether a child is *secure* or *insecure*, the two attachment conditions it is possible to observe in an individual. The results show that the best results were obtained through the combination of decisions made at the level of individual modalities and corresponds to an accuracy of 71.2% (F1 Score 62.4%). The multimodal approach outperforms all unimodal recognizers to a statistically significant extent.

According to John Bowlby, originator of the Attachment Theory, "*[...] the younger the subject the more likely are his behaviour and his mental state to be the two sides of a single coin*" [4]. However, the results of this work seem to contradict such an observation and show that the proposed approach, based on the consistency between observable behaviour and attachment condition, tends to perform better over children of at least 7 years of age (the age range corresponding to levels P3 and P4 at primary school). In particular, the results show that the accuracy of the proposed approach is 76.9% (F1 score 67.0%) for levels P3 and P4, while it is 66.8% (F1 Score 58.8%) for P1 and P2. One possible explanation is that the MCAST administration system used in the experiments (see Section 3) requires children to undergo the assessment without the assistance of an adult. In such a condition, older children might be more likely to provide reliable information because they tend to be more autonomous and capable to perform a task without help.

The main implication of the observations above is that not all children for which the MCAST is designed can equally benefit from the automation of attachment assessment. This is important because insecure attachment should be detected as early as possible to avoid the negative consequences described in Section 1. This opens two possible research avenues, namely the improvement of the recognition approach (e.g., by taking into account further modalities or by improving methodologies for behaviour analysis) and the adaptation of the administration system (e.g., by adding a function capable to assist children or to attract the attention of an adult that can provide help).

In addition to the above, another possibility is to associate a confidence measure to the decisions of the system. Such a technique is commonly used to discriminate between people for which an approach can be trusted and people for which it cannot. For example, in the case of depression, confidence measures were shown to reduce the workload of doctors by roughly two thirds, while still maintaining a performance comparable to an average General Practitioner [1]. The main advantage of a confidence measure is that it does not require one to exclude certain participants a-priori (e.g., younger children according to the considerations above) and it allows an approach to identify as many cases as possible.

One of the main reasons for automating the administration of psychiatric tests is to make the detection of mental health issues more efficient, i.e., to reduce the amount of time needed to verify whether a person is affected by a problem or not. In the case of MCAST, the bottleneck is the time needed for children to complete the different story stems. The experiments suggest that longer recordings do not correspond to higher chances of correct classification. However, this might be the case because all children have been talking for a substantial amount of time (Section 5 shows that the smallest amount of material corresponding to one child is close to 3 minutes). Therefore, one possible direction for future work is to estimate the minimum amount of time needed to achieve a satisfactory performance. Once such a minimum is known, it is possible to investigate whether the MCAST can be modified in such a way that children do not need to speak more than necessary. This would help to further improve the efficiency of the process.

To the best of our knowledge, this article proposes the first multimodal approach aimed at recognizing attachment in children of age between 5 and 9 (previous multimodal approaches were designed for adults [30]). While not being a pathology, insecure attachment leads to lower quality of life [23] and, in some cases, to problems as serious as antisocial behaviour and suicidal tendencies (see Section 1). Some people affected by these extreme consequences cost to society ten times more than an average individual [38]. Therefore,

addressing the attachment recognition problem promises to limit such a major societal burden. In addition, according to the Gartner Group, a major strategic consulting firm, mental health is likely to become one of the most important application areas for Artificial Intelligence (https://www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology/). Correspondingly, attachment recognition can become an important benchmark to assess the effectiveness of AI in supporting the work of psychiatrists.

ACKNOWLEDGMENTS

The research leading to these results was supported by UKRI and EPSRC through grants EP/S02266X/1 and EP/N035305/1, respectively.

REFERENCES

[1] N. Aloshban, A. Esposito, and A. Vinciarelli. 2021. What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech. *Cognitive Computation (to appear)* (2021).
[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 1–10.
[3] C.M. Bishop. 2006. *Pattern recognition and machine learning*. Springer Verlag.
[4] J. Bowlby. 1969. *Attachment and Loss*. The Hogarth Press and the Institute of Psycho-Analysis. 1–401 pages.
[5] C.L. Breazeal. 2004. *Designing Sociable Robots*. MIT press.
[6] A.W. Collins (Ed.). 1984. *Cognitive development in school-age children: Conclusions and new directions*. National Academy Press.
[7] N.L. Collins and L.M. Allard. 2001. Cognitive representations of attachment: The content and function of working models. In *Blackwell Handbook of Social Psychology: Interpersonal Processes*, G.J.O. Fletcher and M.S. Clark (Eds.). Wiley Online Library, 60–85.
[8] G. Doherty, D. Coyle, and M. Matthews. 2010. Design and evaluation guidelines for mental health technologies. *Interacting with Computers* 22, 4 (2010), 243–252.
[9] M. Dong, W.H. Giles, V.J. Felitti, S.R. Dube, J.E. Williams, D.P. Chapman, and R.F. Anda. 2004. Insights into causal pathways for ischemic heart disease: adverse childhood experiences study. *Circulation* 110, 13 (2004), 1761–1766.
[10] M. Esposito, L. Parisi, B. Gallai, R. Marotta, A. Di Dona, S.M. Lavano, M. Roccella, and M. Carotenuto. 2013. Attachment styles in children affected by migraine without aura. *Neuropsychiatric Disease and Treatment* 9 (2013), 1513–1519.
[11] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent developments in OpenSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the ACM International Conference on Multimedia*. 835–838.
[12] F. Eyben, M. Woellmer, and B. Schueller. 2010. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of ACM International Conference on Multimedia*. 1459–1462.
[13] N. Freed, J. Qi, A. Setapen, C. Breazeal, L. Buechley, and H. Raffle. 2011. Sticking Together: Handcrafting Personalized Communication Interfaces. In *Proceedings of the ACM International Conference on Interaction Design and Children*. 238–241.
[14] J. Green, C. Stanley, R. Goldwyn, and V. Smith. 2016. *Coding Manualfor the Manchester Child Attachment Story Task* (version 29 ed.). University of Manchester.
[15] J. Green, C. Stanley, V. Smith, and R. Goldwyn. 2000. A new method of evaluating attachment representations in young school-age children: The Manchester Child Attachment Story Task. *Attachment & Human Development* 2, 1 (2000), 48–70.
[16] C. Harbig, M. Burton, M. Melkumyan, L. Zhang, and J. Choi. 2011. SignBright: A Storytelling Application to Connect Deaf Children and Hearing Parents. In *Proceedings of CHI*. 977–982.
[17] D.C. Herath, C. Kroos, C. Stevens, and D. Burnham. 2013. Adopt-a-robot: A Story of Attachment. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 135–136.
[18] A. Hiolle, K.A. Bard, and L. Canamero. 2009. Assessing human reactions to different robot attachment profiles. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*. 251–256.
[19] H. Ishihara, Y. Yoshikawa, and M. Asada. 2011. Realistic child robot "Affetto" for understanding the caregiver-child attachment relationship that guides the child development. In *Proceedings of the IEEE International Conference on Development and Learning*, Vol. 2. 1–5.
[20] J. Kaye, M. Nelimarkka, R. Kauppinen, S. Vartiainen, and P. Isosomppi. 2011. Mobile Family Interaction: How to Use Mobile Technology to Bring Trust, Safety and Wellbeing into Families. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*. 721–724.
[21] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239.
[22] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč. 2016. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing* 10, 2 (2016), 242–255.
[23] P. Lovenheim. 2018. *The Attachment Effect*. Tarcher Perigee.
[24] A. Meschtscherjakov. 2009. Mobile Attachment: Emotional Attachment Towards Mobile Devices and Services. In *Proceedings of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services*. 102:1–102:1.
[25] A. Meschtscherjakov, D. Wilfinger, and M. Tscheligi. 2014. Mobile attachment causes and consequences for emotional bonding with mobile phones. In *Proceedings of CHI*. 2317–2326.
[26] E. Moss, C. Cyr, and K. Dubois-Comtois. 2004. Attachment at early school age and developmental risk: examining family contexts and behavior problems of controlling-caregiving, controlling-punitive, and behaviorally disorganized children. *Developmental Psychology* 40, 4 (2004), 519–532.
[27] W. Odom, J. Pierce, E. Stolterman, and E. Blevis. 2009. Understanding why we preserve some things and discard others in the context of interaction design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1053–1062.
[28] M. Oyama-Higa, J. Tsujino, and M. Tanabiki. 2006. Does a Mother's Attachment to Her Child Affect Biological Information provided by the Child? -Chaos analysis of fingertip pulse waves of children. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3. 2030–2034.
[29] C.J. Packard, V. Bezlyak, J.S. McLean, G.D. Batty, I. Ford, H. Burns, J. Cavanagh, K.A. Deans, M. Henderson, and A. McGinty. 2011. Early life socioeconomic adversity is associated in adult life with chronic inflammation, carotid atherosclerosis, poorer lung function and decreased cognitive performance: a cross-sectional, population-based study. *BMC Public Health* 11, 1 (2011), 42.
[30] F. Parra, S. Scherer, Y. Benezeth, P. Tsvetanova, and S. Tereno. 2021. Development and cross-cultural evaluation of a scoring algorithm for the Biometric Attachment Test: Overcoming the challenges of multimodal fusion with" small data". *IEEE Transactions on Affective Computing (to appear)* (2021).
[31] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026* (2013).
[32] R. Pascanu, T. Mikolov, and Y. Bengio. 2013. On the difficulty of training Recurrent Neural Networks. In *Proceedings of the International Conference on Machine Learning*. 1310–1318.
[33] R. Ranawana and V. Palade. 2006. Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems* 3, 1 (2006), 35–61.
[34] C. Remy, S. Gegenbauer, and E.M. Huang. 2015. Bridging the Theory-Practice Gap: Lessons and Challenges of Applying the Attachment Framework for Sustainable HCI Design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1305–1314.
[35] G. Roffo, D.-B. Vo, M. Tayarani, M. Rooksby, A. Sorrentino, S. Di Folco, H. Minnis, S. Brewster, and A. Vinciarelli. 2019. Automating the Administration and Analysis of Psychiatric Tests: The Case of Attachment in School Age Children. In *Proceedings of CHI*.
[36] B. Schuller and A. Batliner. 2014. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.
[37] B. Schuller, S. Steidl, and A. Batliner. 2009. The Interspeech 2009 Emotion Challenge. In *Proceedings of Interspeech*.
[38] S. Scott, M. Knapp, J. Henderson, and B. Maughan. 2001. Financial cost of social exclusion: follow up study of antisocial children into adulthood. *British Medical Journal* 323, 7306 (2001), 191.
[39] S.L. Toth, A. Maughan, M. Manly, J.T.and Spagnola, and D. Cicchetti. 2002. The relative efficacy of two interventions in altering maltreated preschool children's representational models: Implications for attachment theory. *Development and Psychopathology* 14, 4 (2002), 877–908.
[40] J. Tsujino, M. Oyama-Higa, and M. Tanabiki. 2008. Measurement of ear pulse waves in children: Effect of facing another child and relationship to an action index. In *2008 IEEE International Conference on Systems, Man and Cybernetics*. 2972–2976.
[41] A. Vinciarelli, M. Pantic, and H. Bourlard. 2009. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.
[42] D.B. Vo, S. Brewster, and A. Vinciarelli. 2020. Did the Children Behave? Investigating the Relationship Between Attachment Condition and Child Computer Interaction. In *Proceedings of the International Conference on Multimodal Interaction*. 88–96.
[43] D. Wilkins, D. Shemmings, and Y. Shemmings. 2015. *Attachment*. Palgrave.
[44] P. Wilson, P. Bradshaw, S. Tipping, G. Der, and H. Minnis. 2013. What predicts persistent early conduct problems? Evidence from the Growing Up in Scotland cohort. *Journal of Epidemiology and Community Health* 67 (2013), 76–80.