ATTACHMENT RECOGNITION IN SCHOOL-AGE CHILDREN: A MULTIMODAL APPROACH BASED ON LANGUAGE AND PARALANGUAGE ANALYSIS

Huda Alsofyani and Alessandro Vinciarelli

University of Glasgow (UK)

ABSTRACT

Attachment is the psychological construct accounting for whether parents address effectively physical and emotional needs of their children or not. The approach proposed in this work recognizes whether a child is secure or insecure, the two major attachment conditions an individual can belong to. The approach is based on the combination of language and paralanguage, what children say and how they say it. The experiments involved 104 children of age between 5 and 9 that were recorded while undergoing the Manchester Attachment Story Task, one of the main psychometric instruments child psychiatrists use to assess the attachment condition of children. The results show that it is possible to achieve an accuracy of up 74.6% (F1 Score 66.7%), meaning that the approach correctly identifies the attachment condition of a child three times out of four, on average.

Index Terms— Social Signal Processing, Language, Paralanguage, Attachment, Multimodal Analysis

1. INTRODUCTION

Attachment is 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*" [1]. For this reason, the attachment condition of an individual is said to be either *secure* or *insecure*. The goal of this work is to show that it is possible to infer the attachment condition of a child through the joint analysis of language and paralanguage, i.e., of *what* children say and *how* they say it. The main reason why such a task is important is that insecure attachment, if not detected and addressed early enough, increases significantly the chances of an individual to experience major issues in adult life, including anti-social behavior [2] and coronary pathologies [3].

The experiments of this work involved 104 children between 5 and 9 years old randomly recruited in the primary schools of Glasgow (see Section 2). The children were recorded while undergoing the *Manchester Child Attachment Story Task* (MCAST) [4], one of the tests that child psychiatrists apply more commonly to assess the attachment condition of their patients. The key-aspect of the MCAST is that participants have to tell stories about every day interactions between children and their parents. The main assumption underlying the MCAST is that children in different attachment conditions will tend to tell different stories and in a different way. For such a reason, the approach proposed in this work is based on the multimodal analysis of language and paralanguage.

To the best of our knowledge, this is one of the first works addressing the problem of attachment recognition in children. Earlier attempts to address the same task were based on gestures that children display while playing with dolls [5], on paralanguage [6] (including its combination with facial expressions [7]) or on blood pressure measured through ear pulse waves (this latter work does not propose an automatic recognition approach, but it shows that attachment leaves physical traces that can be detected and analyzed) [8]. Another work takes into account multiple behavioral channels (language, paralanguage facial expression and physiological signals), but it proposes experiments on adults and not on children. The remaining computing works targeted either technologies aimed at developing positive attachment relationship between children and parents (see, e.g., [9, 10]) or at fostering attachment relationships between users and technologies (see, e.g., [11, 12]).

The rest of this article is organized as follows: Section 2 describes MCAST and the data used in the experiments, Section 3 shows the proposed approach, Section 4 reports on experiments and results, and the final Section 5 draws some conclusions.

2. ATTACHMENT ASSESSMENT AND DATA

The MCAST (see Section 1) [4] is one of the tests psychiatrists apply more commonly to assess the attachment condition of children. During its administration, participants are provided with five *story stems*[1] that they are then asked to con-

The work was supported by UKRI and EPSRC through grants EP/N035305/1, EP/M025055/1 and EP/S02266X/1.

[1]The stems describe the interaction between children and mothers in everyday life and are called *Breakfast* (the mother prepares the breakfast for the child when this latter wakes up), *Nightmare* (the child calls her mother after waking up during a nightmare), *Hopscotch* (the child has an accident whil

tinue and conclude. The way participants continue the stems, corresponding to both content of what they say and nonverbal behaviour displayed during the task, allows the administrators to assess the attchment condition of participants.

During the experiments of this work, the MCAST was delivered with the *School Attachment Monitor* [5], an automatic system that guides participants through the steps above and records them while they enact the story stems with the dolls. Such an approach is in line with the clinical practice of child psychiatrists that typically record the test administration and then use the recordings to analyze the behaviour of the test participants in detail. In this work, the recordings were analyzed by a pool of four assessors trained at the MCAST rating course [13]. Every child was analyzed independently by two assessors and rated as as *secure* or *insecure*. In case of agreement, the assessment was confirmed, while in case of disagreement, the child was discussed by the four assessors together until consensus was reached (an approach typical of clinical practice where difficult and ambigous cases are discussed collegially). In this way, it was possible to rate all children involved in the experiments (see below) as either secure or insecure.

In total, the experiments involved 104 children of age between 5 and 9 randomly recruited in the primary schools of Glasgow (UK). The recruitment was performed according to the ethical guidelines of the British Psychological Association and children were involved only upon written authorization of their parents or caregivers. In addition, the children were free to leave and interrupt the experiment at any moment they wished. Table 1 shows the distribution across school levels (Primary 1 to Primary 4), genders and attachment conditions. According to a χ-square test with confidence level 99%, the attachment distribution in Table 1 lies within a statistical fluctuation with respect to the distribution observed in the general population [14, 15]. The total length of the recordings is 18 hours, 30 minutes and 34 seconds (the average is 640.7 seconds per child). Correspondingly, the automatic transcriptions obtained with *Sonix* (`http://sonix.ai`) include 19,072 words, an average of 183.4 words per child.

Level	P1 (5-6)	P2 (6-7)	P3 (7-8)	P4 (8-9)
Female	9 (8.6%)	22 (20.9%)	15 (15.2%)	11 (10.5%)
Male	10 (9.5%)	18 (17.1%)	14 (13.3%)	5 (4.9%)
Secure	9 (8.6%)	22 (20.9%)	18 (17.1%)	10 (9.6%)
Insecure	10 (9.5%)	18 (17.1%)	11 (11.4%)	6 (5.8%)
Total	19 (18.1%)	40 (38.1%)	30 (28.6%)	16 (15.2%)

Table 1. The table shows the distribution of gender and attachment condition across the primary school levels, *Primary 1* (P1) to *Primary 4* (P4). For every level, the header shows the corresponding age-range between parentheses.

3. THE APPROACH

The proposed multimodal approach builds upon two unimodal recognizers, one based on language and one based on paralanguage. The combination of their outcomes is performed through Weighted Averaging (WA), i.e., by estimating the probability the unimodal systems attribute to the two classes (*secure* and *insecure*) and by then assigning a child to the class that corresponds to the maximum average probability.

The language-based unimodal approach includes three different steps: *preprocessing*, *classification* and *aggregation* of the decisions made at the level of individual story stems. The first step eliminates punctuation, non-alphabetic characters and numbers from the transcriptions. The words are then stemmed, meaning that all morphological variants of the same word (e.g. singular and plural of the same noun) are converted into the same term. Finally, all non-content words (articles, prepositions, etc.) are removed with *Natural Language Toolkit* (`https://www.nltk.org`), a publicly available package for text processing. The resulting sequence of terms is tokenized (every word is replaced with a 1 out of K representation) and padded to a common length L (sequences longer than L are truncated).

The sequence of vectors resulting from the preprocessing is fed to a deep network for the classification step. The first layer of the network performs word embedding [16], i.e. it converts the 1 out of K representations above into lower-dimensional vectors. The following three layers perform 1-D convolution, max pooling and dropout, respectively. The max pooling layer aims at preserving the most useful features resulting from the 1-D convolution, while the dropout layer enhances the generalization properties of the approach. Finally a softmax layer performs the classification, i.e., it estimates the probability of a given child being insecure.

Given that the MCAST includes five story stems, there are 5 transcriptions per child. The approach above is trained individually over the different stems and the decision made individually for each of them are aggregated through Weighted Averaging (see the beginning of this Section).

In the case of paralanguage, the unimodal approach (see [6] for a full description) converts the speech signal into a sequence of vectors $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ extracted from non-overlapping analysis windows of length 33 ms (N is the total number of vectors in X). The vectors were extracted with OpenSmile [17] and their $D = 32$ features were originally designed for a benchmarking campaign on emotion recognition [18]. Since then, the features were shown to account for the main paralinguistic aspects of speech and to be effective in the recognition of a wide spectrum of social and psychological phenomena.

playing hopscotch with other children), *Tummyache* (the child asks comfort while experiencig tummyache) and *Shopping* (the child looses contact with her mother during a visit to a shopping mall).

The vector sequences X were split into subsequences of $L = 128$ vectors and each of them was fed to two *stacked* Recurrent Neural Networks (RNN) [19], where stacked means that the sequence of the hidden states of the first is given as input to the second. It was then this latter to estimate the probability of a recording belonging to class insecure. Given that there were multiple subsequences per recording (the length of X was typically greater than the length L of the subsequences), the final decision about a recording was made through a *majority vote*, meaning that the recording was assigned to the class its subsequences were most frequently assigned to. Like in the case of the language based approach, a different model was trained for each story stem and the decisions made at the level of individual stems were aggregated through Weighted Averaging (see Section 4 for more details).

4. EXPERIMENTS AND RESULTS

The experiment participants were randomly split into $k = 10$ disjoint groups. Correspondingly, the data were split into $k = 10$ folds, each containing all the recordings of the participants in a given group. In this way, it was possible to perform the experiments according to a k-fold protocol and to ensure that children were never represented in both training and test material. This made the experiments *person independent*, i.e., it ensured that the approach recognized the attachment condition of the participants and and not their identity.

All hyper-parameters of the models described in Section 3 were set to standard values common in the literature (no attempt was made to find better values through cross-validation). For the language-based model, all text sequences were padded to length 100, with the word embedding dimension set to 300. the number of filters of the convolution layer is 64 and three different kernel sizes were tested through crossvalidation (3, 5 and 7) with 5 giving the best performance. The dropout rate was set to 0.2. The training was performed with a mini-batch strategy to limit computational issues. Each mini batch includes ten input sequences and the number of training epochs was $T = 20$.

For the RNNs of the paralanguage-based approach, the number of training epochs was set to $T = 50$, the dimension of the hidden states to $D = 70$, the learning rate to $\beta = 10^{-3}$, and the length of the input sequences to $L = 128$. Overfitting risks were reduced by applying L2 regularization with parameter $\lambda = 10^{-2}$. Computational issues were addressed through a mini-batch strategy [20], i.e. by training the RNNs over different subsets of the data, each including $B = 512$ sequences. All experiments (using both language and paralanguage) were performed $R = 10$ times because the training process involves a random initialization step that can lead to different results. Correspondingly, all performance metrics are reported in terms of average and standard deviation over the R repetitions.

Since the different story stems tend to elicit different reactions, the networks were trained separately over the material corresponding to each of them, thus resulting into 5 different models (one per stem). The main advantage of such an approach is that the different reactions can be a source of *diversity*, i.e., the tendency of different classifiers to make different mistakes over different samples [21]. Such a property is important because it can help an ensemble of classifiers to improve over its best member and, for this reason, Table 2 shows the results obtained not only at the level of individual stems, but also at the level of the aggregation performed through weighted averaging (see Section 3).

Story	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Speech				
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
Transcriptions				
Breakfast	71.2±2.0	69.3±3.1	59.5±2.2	64.0±2.2
Nightmare	65.1±2.9	62.3±4.6	49.3±4.2	55.0±4.0
Tummyache	64.4±2.3	62.7±4.1	43.4±2.5	51.3±2.9
Hopscotch	70.8±2.0	70.6±2.8	54.1±3.7	61.2±3.0
Shop. Mall	69.5±1.9	66.7±2.8	57.3±3.0	61.6±2.5
All (WA)	71.3±2.1	71.7±3.0	55.8±3.5	62.7±3.1
Multimodal				
Breakfast	71.5±2.0	70.4±2.8	58.1±2.7	63.6±2.5
Nightmare	68.8±2.2	69.3±4.4	50.2±4.7	58.1±3.5
Tummyache	68.3±2.3	68.6±4.2	49.3±4.2	57.3±3.4
Hopscotch	72.7±1.7	74.8±3.8	54.8±3.9	63.1±2.5
Shop. Mall	69.4±1.9	67.5±2.3	54.8±3.9	60.4±3.1
All (WA)	74.6±2.0	77.5±4.7	58.7±2.1	66.7±2.1
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 models have been initialized differently). The Random classifier assigns samples to classes according to a-priori probabilities.

Overall, Table 2 shows that all models are better than chance, i.e., they improve to a statistically significant extent the performance of a random classifier assigning a sample to class c with probability $p(c)$ corresponding to the prior of c ($p < 0.01$ according to a t-test). However, the performance changes from one story stem to the other for individual modalities and for their combination. In the case of paralanguage, the difference between highest accuracy (65.8% for Breakfast) and bottom two accuracies (61.0% and 60.1% for Nightmare and Tummyache, respectively) is statistically significant ($p < 0.01$ according to a two-tailed t-test in both cases). The same applies to language, where the highest accuracy (71.2% for *Breakfast*) is statistically significantly higher

Level	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
P1	68.4±5.5	78.2±9.4	56.0±5.2	65.1±6.0
P2	77.0±2.8	82.6±5.0	62.2±5.1	70.8±3.8
P3	76.9±2.3	79.3±6.8	53.6±2.9	63.8±2.9
P4	71.9±9.0	65.8±18.7	61.7±8.1	62.8±10.2

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

than the bottom three (69.5% for Shopping Mall, 65.1% for Nightmare and 64.4% for Tummyache). The same for the multimodal combination, where the best performing stem is Hopscotch (72.7%) and the least performing ones are Shopping Mall, Nightmare and Tummyache (69.4%, 68.8% and 68.3%, respectively).

Overall, the results above confirm the assumption that different stems tend to elicit attachment relevant behaviors to a different extent. The question that remains open is whether this is a source of diversity, i.e., whether models trained over different story stems tend to make different mistakes over different children. This seems to be the case for paralanguage, where the aggregation of individual story outcomes leads to a statistically significant improvement over the best performing stem ($p < 0.05$ according to a two-tailed t-test), thus suggesting that children change the way they tell the stories depending on the stem. However, it seems not to be the case for language, where the aggregation of the outcomes does not improve over the best stem. This suggests that the traces of attachment in language tend to remain the same across the different stems.

Another potential source of diversity comes from the presence of two modalities. According to the results of Table 2, the multimodal combination of paralanguage and language improves, to a statistically significant extent, over both individual modalities ($p < 0.001$ according to a two-tailed t-test). This seems to suggest that paralanguage and language carry, at least to a certain extent, complementary information that can help the two modalities to mutually correct each other. In particular, the multimodal approach involving the weighted average reaches an accuracy of 74.6% and, therefore, it correctly recognizes the attachment condition of children three times out of four, on average. The corresponding Recall is 58.7% and such a values shows that almost two thirds of the insecure children are correctly identified as such.

One of the main challenges for the proposed approach is that it deals with children of age between 5 and 9 (see Table 1). Such an age range covers widely different developmental stages and this suggests that not all children can be equally effective at participating in the MCAST. For this reason, Table 3 shows the results obtained for different school levels with the multimodal combination of language and paralanguage (after WA). The difference between level P1 and levels P2 and P3 is statistically significant ($p < 0.001$ according to a two-tailed t-test), but the same does not apply to the difference between P1 and P4. In this respect, while the performance seems to increase when passing from P1 to P2 and P3, it is not clear whether such a pattern remains when passing from P1 to P4. One possible explanation of the results is that P1 children might be more in difficulty in dealing with the MCAST, while those at level P4 start being too old to feel comfortable at playing with dolls. However, it cannot be excluded that the lower accuracy at level 4 is simply an artefact due to the limited number of P4 participants (the variance is higher than in the other cases).

5. CONCLUSIONS

This article presents experiments on attachment recognition in school-age children. To the best of our knowledge, this is the first approach that addresses the problem through the multimodal analysis of language and paralanguage, what children say and how they say it. The results show that the proposed approach can reach an accuracy of up to 74.6% (F1 Score 66.7%) and that the multimodal approach improves over both unimodal approaches, thus showing that the two behavioural channels tend to carry complementary information. In addition, the experiments seem to suggest that the performance of the approach tends to improve with the age of the children. However, the limited number of children at the top of the age-range does not allow one to reach conclusive results about this point.

Overall, the results seem to suggest that it is possible to identify roughly two thirds of the insecure children. This is important because, while not being a pathology, insecure attachment increases the chances of experiencing negative issues [2, 3] and, overall, it can reduce significantly the quality of life [1]. Such negative effects can be attenuated if insecure attachment is detected early enough. However, current methodologies for attachment assessment are expensive and time-consuming and most insecure children are not identified as such. In this respect, automatic methodologies like those proposed in this work have the potential to allow large-screenings of the population that can lead to the identification of a large fraction of the insecure children. A major progress with respect to the current situation.

Future work will address two main avenues, namely the inclusion of further behavioral modalities in the approach (e.g., facial expressions) and the identification of attachment markers. In the first case, the main challenge will be the development of representation strategies capable to effectively model multiple behavioral streams [22]. In the second case, the main challenge will be the identification of approaches capable to provide information about the behavioral cues most likely to lead to the correct classification of the children.

6. REFERENCES

[1] P. Lovenheim, *The Attachment Effect*, Tarcher Perigee, 2018.

[2] P. Wilson, P. Bradshaw, S. Tipping, G. Der, and H. Minnis, "What predicts persistent early conduct problems? Evidence from the growing up in Scotland cohort," *Journal of Epidemiology and Community Health*, vol. 67, pp. 76–80, 2013.

[3] M. Dong, W.H. Giles, V.J. Felitti, S.R. Dube, J.E. Williams, D.P. Chapman, and R.F. Anda, "Insights into causal pathways for ischemic heart disease: adverse childhood experiences study," *Circulation*, vol. 110, no. 13, pp. 1761–1766, 2004.

[4] J. Green, C. Stanley, V. Smith, and R. Goldwyn, "A new method of evaluating attachment representations in young school-age children: The Manchester Child Attachment Story Task," *Attachment & Human Development*, vol. 2, no. 1, pp. 48–70, 2000.

[5] G. Roffo, D.-B. Vo, M/ Tayarani, M. Rooksby, A. Sorrentino, S. Di Folco, H. Minnis, S. Brewster, and A. Vinciarelli, "Automating the administration and analysis of psychiatric tests: The case of attachment in school age children," in *Proceedings of CHI*, 2019.

[6] H. Alsofyani and A. Vinciarelli, "Stacked recurrent neural networks for speech-based inference of attachment condition in school age children," in *Proceedings of Interspeech*, 2021.

[7] H. Alsofyani and A. Vinciarelli, "Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2021.

[8] M. Oyama-Higa, J. Tsujino, and M. Tanabiki, "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*, 2006, vol. 3, pp. 2030–2034.

[9] C. Harbig, M. Burton, M. Melkumyan, L. Zhang, and J. Choi, "SignBright: A storytelling application to connect deaf children and hearing parents," in *Proceedings of CHI*, 2011, pp. 977–982.

[10] A. Hiolle, K.A. Bard, and L. Canamero, "Assessing human reactions to different robot attachment profiles," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 251–256.

[11] D.C. Herath, C. Kroos, C. Stevens, and D. Burnham, "Adopt-a-robot: A story of attachment," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2013, pp. 135–136.

[12] A. Meschtscherjakov, D. Wilfinger, and M. Tscheligi, "Mobile attachment causes and consequences for emotional bonding with mobile phones," in *Proceedings of CHI*, 2014, pp. 2317–2326.

[13] J. Green, C. Stanley, R. Goldwyn, and V. Smith, *Coding Manualfor the Manchester Child Attachment Story Task*, University of Manchester, version 29 edition, 2016.

[14] M. Esposito, L. Parisi, B. Gallai, R. Marotta, A. Di Dona, S.M. Lavano, M. Roccella, and M. Carotenuto, "Attachment styles in children affected by migraine without aura," *Neuropsychiatric Disease and Treatment*, vol. 9, pp. 1513–1519, 2013.

[15] E. Moss, C. Cyr, and K. Dubois-Comtois, "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*, vol. 40, no. 4, pp. 519–532, 2004.

[16] E. Charniak, *Inreoduction to Deep Learning*, MIT Press, 2018.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in OpenSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 835–838.

[18] Björn Schuller, Stefan Steidl, and Anton Batliner, "The Interspeech 2009 Emotion Challenge," in *Proceedings of Interspeech*, 2009.

[19] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.

[20] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč, "Mini-batch semi-stochastic gradient descent in the proximal setting," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 242–255, 2016.

[21] R. Ranawana and V. Palade, "Multi-classifier systems: Review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, vol. 3, no. 1, pp. 35–61, 2006.

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.