INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Stacked Recurrent Neural Networks for Speech-Based Inference of Attachment Condition in School Age Children

Huda Alsofyani, Alessandro Vinciarelli

University of Glasgow, UK
firtsname.lastname@glasgow.ac.uk

Abstract

In Attachment Theory, children that have a positive perception of their parents are said to be secure, while the others are said to be insecure. Once adult, unless identified and supported early enough, insecure children have higher chances to experience major issues (e.g., suicidal tendencies and antisocial behavior). For this reason, this article proposes a speech-based automatic approach for the recognition of attachment in school-age children. The experiments are based on stacked RNNs and have involved 104 children of age between 5 and 9. The accuracy is up to 68.9% (F1 59.6%), meaning that the approach makes the right decision two times out of three, on average. To the best of our knowledge, this is the first work aimed at inferring attachment from speech in school-age children.

Index Terms: Computational paralinguistics, attachment, child speech, Social Signal Processing.

1. Introduction

According to John Bowlby, originator of the Attachment Theory, "*to know that an attachment figure is available and responsive gives [children] a strong and pervasive feeling of security*" [1]. Therefore, attachment can be thought of as 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*" [2]. Correspondingly, the attachment condition of a child is said to be *secure* or *insecure* depending on whether children perceive their attachment figures, typically the parents, in positive or negative terms, respectively [3].

The main reason why the attachment condition is important is that insecure attachment, if not properly addressed during childhood, can have negative consequences in adult life. For example, insecure children are more likely to display antisocial behavior [4] or to develop coronary pathologies [5, 6] once they become adult. Furthermore, attachment shapes beliefs and expectations about relationships in general [7]. As a consequence, insecure individuals tend to be less satisfied with their social life, marriage or professional career because they tend to perceive more negatively friends, romantic partners or colleagues, respectively [2]. For these reasons, this article proposes a speech-based automatic approach for attachment recognition in children of age between 5 and 9.

The proposed approach applies methodologies typical of Social Signal Processing [8] and Computational Paralinguistics [9]. In particular, the approach recognizes whether children are *secure* or *insecure* by analyzing nonverbal aspects of speech. The approach starts by converting input speech signals into sequences of feature vectors extracted at regular time steps. It then feeds such sequences to stacked Recurrent Neural Networks (RNN) [10] trained to distinguish between secure and insecure child speakers. The main reason for focusing on nonverbal aspects is that they were shown to convey reliable information about social and psychological phenomena such as, e.g., emotions [11] and personality traits [12].

The experiments have involved 104 participants (59 secure and 45 insecure) undergoing the *Manchester Child Attachment Story Task* (MCAST) [13], one of the tests child psychiatrists apply most commonly in clinical practice. During such a test, children are recorded while telling stories about everyday interactions between two fictitious characters, a mother and her child (see Section 2 for more details). According to the theory underlying the MCAST, children manifest their attachment condition through the way they tell such stories, a scenario that naturally lends itself to the application of the approach described above. The total duration of the recordings is 7 hours, 1 minute and 24 seconds, corresponding to an average of 240.8 seconds per child. The results show that the attachment recognition accuracy is 68.9% (F1 59.6%), meaning that the approach makes the right decision roughly two times out of three.

To the best of our knowledge, this is the first attempt to automatically infer the attachment condition of children from speech. Experiments similar to those presented in this work were shown in [14], but the focus was on the way children move dolls representing the characters at the core of the MCAST (the results are similar to those obtained in this article). An automatic version of the *Biometric Attachment Test* was proposed in [15], but it was designed for adults and not for children. Such a work focuses on physiological measurements (photoplethysmography), face behavior, paralinguistics and language to predict attachment self-assessment scores (the best Root Mean Square Error is 12.1). Other computing works revolving around attachment target the development of digital artifacts capable to establish long-term relationships with their users [16, 17]. Similarly, the role of attachment in child-robot interaction was explored in [18, 19, 20], with a particular focus on social robots designed to interact with their users like humans. Finally, other works aim at designing technologies that support positive attachment relationships between children and caregivers [21, 22, 23].

The rest of this article is organized as follows: Section 2 provides information about the MCAST and the data collected for the experiments, Section 3 describes the attachment recognition approach, Section 4 reports on experiments and results, and the final Section 5 draws some conclusions.

2. Attachment Assessment and Data

The *Manchester Child Attachment Story Task* (MCAST) [13] is one of the instruments that experts use most commonly to assess the attachment condition of a child. The test is based on five story stems about the interaction between a child and a mother in everyday life: *Breakfast* (the child wakes up in the morning and the mother prepares breakfast), *Nightmare* (the child wakes

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-904

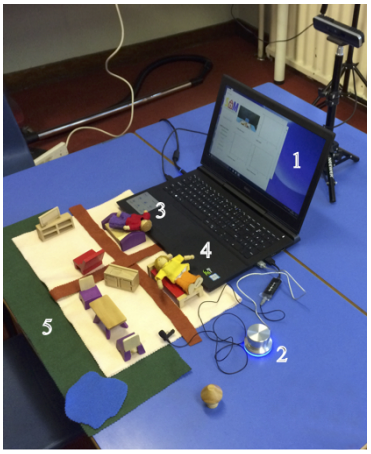

Figure 1: *The picture shows the School Attachment Monitor and how it appears to the children. The computer screen (element 1 in the picture) displays the videos where actors guide the children through the steps of the MCAST, the button (element 2 in the picture) allows the users to signal that they have completed an MCAST step, the dolls (elements 3 and 4) and the toy house (see element 5) allow the users to complete the story stems.*

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) and *Shopping* (the child looses contact with the mother in a shopping mall and tries to re-establish contact with her). After listening to each stem, the participants have to tell how the story continues with the help of two dolls, one that corresponds to the child of the story and the other that corresponds to her mother. The key-assumption underlying the MCAST is that participants in different attachment conditions will tell the stories in a different way. For this reason, it is common clinical practice to record the children undergoing the MCAST so that psychiatrists can analyze the way they tell the stories in full detail.

Figure 1 shows the main elements of the School Attachment Monitor (SAM), the system used for the collection of the data in the experiments. The MCAST administration takes place according to the following protocol:

- *Story stem delivery*: the SAM plays a video in which an actor delivers a story stem and then prompts the participant to represent its continuation with the dolls (see element 1 of the SAM in Figure 1);
- *Story representation*: the participant represents the continuation of the story stem with the help of dolls (see elements 3 and 4 in Figure 1) and play mat (element 5 in Figure 1) while being recorded by the camera of the system and, at the conclusion, presses the "*Finish*" button (element 2 in Figure 1);
- *Iteration*: the system goes back to the first step to deliver another story stem (if the fifth story stem has not been reached) or concludes the test (if the fifth story stem has been reached).

The SAM records the whole administration of the MCAST, but the experiments have been performed only over the segments in which the children actually tell the stories. The reason is that, according to the theory underlying the test, this is when the participants manifest their attachment condition [13].

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.*

Level	P1 (5-6)	P2 (6-7)	P3 (7-8)	P4 (8-9)
Female	9	22	15	11
Male	10	18	14	5
Secure	9	22	18	10
Insecure	10	18	11	6
Total	19	40	29	16

The experiments have involved 104 children randomly recruited in the primary schools of Glasgow, UK (Table 1 shows their distribution across school levels, gender and attachment conditions). The total number of female and male participants is 57 (55.2% of the total) and 48 (44.8% of the total), respectively. The attachment assessment has been performed by a pool of 4 raters that have attended the professional course delivered by the psychiatrists that have elaborated the test [24]. Every child of the corpus has been assessed by two independent raters and, in case of disagreement (less than 10% of the cases), a third rater has been asked to perform an independent assessment aimed at breaking the tie. Such a protocol replicates the practices of the clinicians that have led the collection of the data [14].

According to a χ-square test with confidence level 99%, the corpus distribution over attachment conditions (55.8% of secure and 44.2% of insecure) is within a statistical fluctuation with respect to the distribution observed in the rest of the population [25, 26]. The total length of the resulting recordings collected in the experiments is 7 hours, 1 minute and 24 seconds, corresponding to an average of 240.8 seconds per child. The collection of the data was performed after having received the ethical clearance of the School Authority in Glasgow. Children were involved only upon written authorization of their parents and they were free to interrupt the test at any moment.

3. The Approach

The proposed approach includes three main steps, namely *feature extraction*, *attachment recognition* and *aggregation*. The first step converts the input recordings into sequences of feature vectors, the second assigns such sequences to class *secure* or *insecure*, and the third aggregates the classifications made at the level of individual story stems.

The feature extraction step is performed with OpenSmile [27, 28] over 33 ms long non-overlapping analysis windows. The feature set includes 16 basic features and their respective delta regression coefficients for a total of 32 features. The basic features are *Root mean square of the energy* (1 feature), *Mel Frequency Cepstral Coefficients* (12 features), *Zero Crossing Rate* (1 feature), *Voicing probability* (1 feature) and *Fundamental frequency* (1 feature). The main motivation behind this choice is that the features above were shown to be effective in emotion recognition [29] and are commonly used in the literature to infer social and psychological phenomena from speech. The feature values are smoothed by averaging over three consecutive analysis windows.

For a recording corresponding to an individual story stem, the result of the feature extraction process is a sequence of feature vectors $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. The goal of the attachment recognition step is to assign X to one of the two possible classes

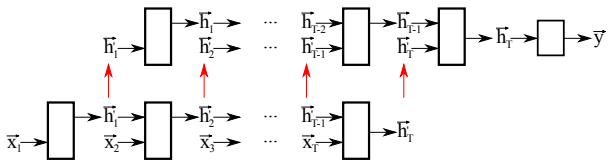

Figure 2: *The figure shows the information flow through the stacked RNNs during the recognition process. The rectangular blocks correspond to hyperbolic tangent layers, while the square block is a softmax layer. Vectors $\boldsymbol{h}_t$ and $\boldsymbol{h}'_t$ are the hidden states of the RNNs, while the $\boldsymbol{x}_t$s are the feature vectors extracted from the speech recordings.*

(secure or insecure). Given the sequential nature of the data, Recurrent Neural Networks (RNN) [30] appear to be a suitable model for recognition. In particular, the experiments were performed using two *stacked* RNNs [10], meaning that the hidden states of the first RNN were fed to the second as input (see Figure 2). The main expectation behind such an architecture is that hidden state sequences of higher networks in the stack tend to account for increasingly higher levels of abstraction.

Training the model of Figure 2 over long sequences gives rise to issues such as vanishing and exploding gradients [31]. Therefore, the input recordings were split into non-overlapping segments that correspond to $L = 128$ vectors each (accounting for 4.25 seconds of speech). Such a length is a tradeoff that has been shown to be sufficient to capture temporal information while avoiding training issues. The classification outcome corresponds to the output y of the final softmax layer when all L vectors of a segment have been fed to the stacked RNNs (see Figure 2). The value of y is an estimate of the probability that sequence X belongs to class *insecure*. When such a probability is above 0.5, the segment is assigned to class *insecure*, otherwise it is assigned to class *secure*.

Every speech recording includes M segments that are individually assigned to one of the two possible classes (see above). Correspondingly, a recording can be classified through a majority vote, i.e., it can be assigned to the class $\hat{c}$ that satisfies the following equation: $\hat{c} = \arg\max_{c \in \{c_1, c_2\}} f(c)$, where $f(c)$ is the fraction of segments assigned to class c in the recording.

Section 2 shows that the MCAST includes five story stems and, therefore, there are five recordings per child. Each of these is classified with the approach presented above and, as a result, there are five classification outcomes per child. The goal of the aggregation step is to combine the decisions made at the level of individual story stems to perform a classification at the level of a child. Such a step can be performed in two ways. The first, referred to as *Story Majority Vote*, is to assign the child to the class her or his stories are more frequently assigned to. The second is to consider the story recordings assigned to a particular class c and to calculate the average value of the fraction $f(c)$ (see above) over them. This will result into two averages $E(c_1)$ and $E(c_2)$ extracted from the story recordings assigned to c_1 and c_2 for a given child, respectively. The classification can then be performed as follows: $c^* = \arg\max_{c \in \{c_1, c_2\}} E(c)$, where c^* is the class the child is assgned to. Such an aggregation approach is referred to as *Weighted Average*.

4. Experiments and Results

The experiments were performed according to a k-fold protocol ($k = 10$). The folds were obtained by randomly assigning the data of every child to one of the folds. Given that the same child was never represented in more than one fold, the protocol is *person-independent*, meaning that the same child is never represented in both training and test set. This ensures that the proposed approach actually recognizes attachment and not children. Section 3 shows that the length of the input sequences was set to $L = 128$ (no cross-validation was performed to find potentially better values). Similarly, other parameters were set to values that are standard in the literature, namely the dimension $D = 70$ of the hidden states, the learning rate equal to 10^{-3} and the number of training epochs $T = 50$. To reduce the risk of overfitting, L2 regularization was applied to both recurrent and kernel weights of the RNN layers with parameter $\lambda = 10^{-2}$. The training was performed with a mini-batch strategy to limit computational issues [32]. This means that the RNNs were trained over subsets of the training set (the mini-batches), each including $B = 512$ training sequences. The mini-batches were disjoint, but their union corresponded to the whole training set. Every recognition experiment was performed $R = 10$ times and, at every repetition, the RNNs were initialized randomly. All results are presented in terms of average and standard deviation over the R repetitions.

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). The acronyms SMV and WA stand for Story Majority Vote and Weighted Average. The Random classifier assigns samples to classes according to a-priori probabilities.*

Story	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
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 (SMV)	66.7±2.0	65.2±2.7	49.3±4.9	56.0±3.9
All (WA)	68.9±2.0	67.8±2.4	53.3±5.0	59.6±3.8
Random	51.0	43.0	43.0	43.0

A different model was trained for each of the five individual story stems, thus resulting into five different models (e.g., the "Breakfast model" was obtained by training the stacked RNNs over the recordings corresponding to the Breakfast story stem). In such a way, the individual models can be used as an ensemble of classifiers [33] and this is an advantage because different stories are likely to elicit attachment relevant behaviors to a different extent. For these reasons, Table 2 shows the results both for the individual story stems and for the aggregation performed through Story Majority Vote or Weighted Average (see end of Section 3). The baseline for comparison is a random classifier that assigns a sample to class i with probability p_i, where p_i is the a-priori probability of class i. According to a single-tailed t-test, the difference with respect to such a random classifier is always statistically significant for Accuracy, Precision and F1 Score ($p < 0.01$ in all cases). In the case of Recall, the difference is statistically significant only for Story Majority Vote and Weighted Average.

The low standard deviations suggest that there is no interplay between results and RNNs' initialization. In the case of Recall the standard deviations are higher, but such a perfor-

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

Level	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
P1	63.7±7.0	67.6±6.7	58.5±12.3	62.3±9.7
P2	72.4±4.0	76.4±3.8	55.6±8.4	64.1±6.8
P3	69.7±3.3	64.8±6.7	44.5±5.8	52.6±5.4
P4	65.0±7.7	54.2±11.2	54.2±9.2	53.8±9.2

mance metric takes into account only the 45 insecure children and, therefore, its value fluctuates more. According to a two-tailed t-test, the difference between highest accuracy (65.8% for Breakfast) and bottom two accuracies (61.0% and 60.1% and 63.0% for Nightmare and Tummyache, respectively) is statistically significant ($p < 0.01$ in both cases). This seems to confirm that some of the story stems are more likely to elicit detectable attachment-related behaviors.

In the case of Recall, there is a statistically significant difference ($p < 0.01$ according to a single tailed t-test) between top and bottom values (48.5% and 41.8% for Shopping Mall and Nightmare, respectively), but not between the others. Recall plays an important role in clinical applications because it accounts for type I errors (insecure participants erroneously classified as secure), those that have the most negative consequences because children that need medical attention do not receive it. One possible explanation behind the result above is that insecure children tend to react more uniformly to the different stories and, therefore, there are no major differences in terms of insecure detection.

The Story Majority Vote does not improve over the best accuracy for an individual story stem (according to a two-tailed t-test). Such a result suggests that the classifiers trained over individual story stems are not *diverse*, i.e., they do not tend to make different mistakes over different children [34]. However, the Weighted Average improves over the best accuracy for an individual stem ($p < 0.01$ according to a two-tailed t-test) and this suggests that the classifiers trained over individual stories, when assigning the majority of the speech segments to the right class, tend to do it to a greater extent. In other words, the fraction $f(c)$ of segments assigned to class c (see end of Section 3) tends to be higher when c is the right class.

Table 1 shows the distribution of the experiment participants across the four primary school levels in Scotland (where the data were collected), from P1 (*Primary 1*) to P4 (*Primary 4*). For this reason, Table 3 presents the results obtained through Weighted Average for children at different school levels. According to a single-tailed t-test, there is no statistically significant difference between the accuracies observed over children of levels P2 and P3. However, the difference is statistically significant between such children and the others ($p < 0.05$ according to a single-tailed t-test). Therefore, the proposed approach seems to work better for children at levels P2 and P3. In particular, when taking into account only children of such levels, the accuracy is 71.1%, the Precision is 71.4%, the Recall is 50.8% and the F1 score is 59.2%. One possible explanation is that such levels account for roughly 66% of the participants (69 out of 104) and, therefore, they correspond to a similar fraction of the training material. In other words, the availability of more children in levels P2 and P3 results into models that are more effective for such levels.

In the case of Recall, the situation is different and, in particular, the only value that is lower than the others to a statistically significant extent is observed for P3 ($p < 0.05$ according to a single-tailed t-test in all cases). This suggests that, in terms of Recall the proposed approach seems to be equally effective across 3 of the 4 levels. Furthermore, it should be considered that Recall takes into account only insecure children (45 out of 104 children) and, therefore, the lower performance of P3 might depend on a fluctuation due to the limited number of children at such a level (11 out of the 45 insecure). This is important because it means that it is possible to identify insecure children with roughly the same level of performance across at least 3 school levels.

5. Conclusions

To the best of our knowledge, this work proposes the first speech-based approach for attachment recognition in children. The results show that it is possible to achieve an accuracy of up to 68.9%, corresponding to an F1 score of 59.6%. The main motivation behind the work is that, according to the guidelines of the UK National Collaborating Centre for Mental Health, "*attachment difficulties [...] place a considerable financial burden on health, social services, criminal justice and society as a whole*" [35]. Therefore, it is necessary to "*develop reliable and valid screening assessment tools for attachment [...] that can be made available and used in routine health and social care*" [35]. In particular, automatic approaches for attachment assessment can allow large-scale screenings of the population that, at the moment, are not possible because traditional, non-automatic assessment tests are too time-consuming.

According to the observations of child psychiatry, "*[...] the younger the subject the more likely are his behavior and his mental state to be the two sides of a single coin*" [36]. For this reason, the proposed approach is based on methodologies typical of Social Signal Processing and Computational Paralinguistics, two computing domains focusing on the analysis of nonverbal behavior. Overall, the results appear to confirm that attachment, like any other social and psychological phenomena, leaves traces in terms of *honest* nonverbal behavioral cues, i.e., cues that convey reliable information about the actual inner state of an individual [8].

One of the main characteristics of the proposed approach is the aggregation of decisions made at the level of individual story stems. The reason is that, according to the theory underlying the MCAST, different stems elicit attachment-relevant reactions to a different extent in different children. In this respect, the use of five stems aims at ensuring that, at least for one stem, every child manifests her or his condition with sufficient evidence. In order to increase the diversity across models trained over individual stems, future work will focus on the inclusion of different modalities, i.e., of different behavioral channels through which attachment can be expressed (e.g., facial expressions, language or gestures). The main motivation is that new modalities can inject diversity, i.e., the tendency to make different mistakes over different children. This is the main property that can help an ensemble of classifiers to improve its performance [34].

6. Acknowledgements

This work was supported by UK Research and Innovation and Engineering and Physical Sciences Research Council through the projects "Socially Competent Robots" (EP/N035305/1), "School Attachment Monitor" (EP/M025055/1) and "UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents" (EP/S02266X/1).

7. References

[1] J. Bowlby, *A secure base*. Basic Books, 1988.

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

[3] D. Wilkins, D. Shemmings, and Y. Shemmings, *Attachment*. Palgrave, 2015.

[4] 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.

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

[6] C. Packard, V. Bezlyak, J. McLean, G. Batty, I. Ford, H. Burns, J. Cavanagh, K. Deans, M. Henderson, and A. McGinty, "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*, vol. 11, no. 1, p. 42, 2011.

[7] N. Collins and L. Allard, "Cognitive representations of attachment: The content and function of working models," in *Blackwell Handbook of Social Psychology: Interpersonal Processes*, G. Fletcher and M. Clark, Eds. Wiley Online Library, 2001, pp. 60–85.

[8] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[9] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2014.

[10] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.

[11] M. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[12] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[13] 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.

[14] 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.

[15] F. Parra, S. Scherer, Y. Benezeth, P. Tsvetanova, and S. Tereno, "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.

[16] A. Meschtscherjakov, "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*, 2009, pp. 102:1–102:1.

[17] 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.

[18] H. Ishihara, Y. Yoshikawa, and M. Asada, "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, 2011, pp. 1–5.

[19] D. 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.

[20] A. Hiolle, K. 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.

[21] N. Freed, J. Qi, A. Setapen, C. Breazeal, L. Buechley, and H. Raffle, "Sticking together: Handcrafting personalized communication interfaces," in *Proceedings of the ACM International Conference on Interaction Design and Children*, 2011, pp. 238–241.

[22] J. Kaye, M. Nelimarkka, R. Kauppinen, S. Vartiainen, and P. Isosomppi, "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*, 2011, pp. 721–724.

[23] 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.

[24] J. Green, C. Stanley, R. Goldwyn, and V. Smith, *Coding Manual for the Manchester Child Attachment Story Task*, version 29 ed., University of Manchester, 2016.

[25] M. Esposito, L. Parisi, B. Gallai, R. Marotta, A. Di Dona, S. 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.

[26] 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.

[27] F. Eyben, M. Woellmer, and B. Schueller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[28] 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.

[29] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proceedings of Interspeech*, 2009.

[30] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*. Elsevier, 1997, vol. 121, pp. 471–495.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training Recurrent Neural Networks," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1310–1318.

[32] 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.

[33] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[34] 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.

[35] "AA.VV.", "Children's attachment," National Collaborating Centre for Mental Health, Tech. Rep., 2015.

[36] J. Bowlby, *Attachment and Loss*. The Hogarth Press and the Institute of Psycho-Analysis, 1969.