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

Huda Alsofyani and Alessandro vinciarelli, University of Glasgow

Paper #4270

Introduction

- Attachment is the psychological construct accounting for whether parents address effectively physical and emotional needs of their children or not.
- The attachment condition of an individual is said to be either secure or insecure.
- Insecure attachment, if not addressed properly, increases significantly the chances to experience major issues in adult life, including anti-social behavior and coronary pathologies.
- This work shows that it is possible to infer the attachment condition of a child through the joint analysis of language and paralanguage (what children say and how they say it).

Data

Video recordings of 104 children of age 5-9 years (59 secure and 45 insecure) undergoing the Manchester Child Attachment Story Task (MCAST). In this task, children are asked to listen and complete five different attachment-related stories: Breakfast (BF), Nightmare (NM), Tummyache (TA), Hopscotch (HS), and Shopping mall (SM). The transcriptions are obtained automatically with Sonix (http://sonix. ai).

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 Multimodal Approach

The proposed approach builds upon two modal recognisers: one for Language and the other for paralanguage. The outcomes are combined 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.

Language-Based Approach

This approach consists of three main steps: preprocessing, classification, and aggregation.

- Preprocessing: Eliminates punctuation, non-alphabetic characters and numbers, stemming, stop-words removing. Subsequently, the resulting term sequences are tokenised and padded to the same length L.
- Classification: A deep network consists of: an embedding layer to convert the 1 out of K representations above (tokenised vocabulary) into lower-dimensional vectors. This is followed by three layers to perform 1-D convolution, max pooling, and dropout, respectively. Finally, a softmax layer performs the classification.
- Aggregation: The approach is trained individually over the different 5 stems and the decision is made individually for each of them. These decisions are aggregated through the WA method described above.

Paralanguage-Based Approach

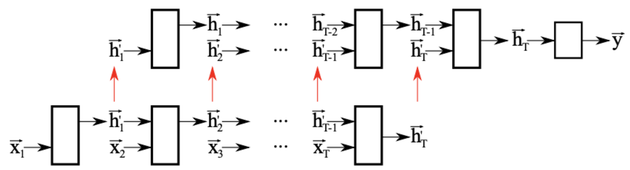
This approach consists of three main steps: feature extraction, classification, and aggregation.

- Feature Extraction: The features are extracted using OpenSmile[1] over 33 ms long non-overlapping analysis windows. The feature set consists of 16 basic features and their respective delta regression coefficients, resulting in 32 features that were shown to be effective in emotion recognition. The result of the feature extraction is a sequence of feature vectors $X = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T})$. The basic features are as follows:

Root mean square of the energy	1 feature
Mel Frequency Cepstral Coefficients	12 features
Zero Crossing Rate	1 feature
Voicing probability	1 feature
Fundamental frequency	1 feature

The feature values are smoothed by averaging over three consecutive analysis windows.

- Classification: As the data denote sequences, a Recurrent Neural Networks (RNN) is used for the recognition step. In particular, two stacked RNN layers are used where the hidden states of the first layer are fed as input to the second layer.



The input sequences are split into non-overlapping segments to vectors of length L=128 to avoid computational issues such as vanishing and exploding gradients. Each segment is assigned individually to one of the classes and the full recording is assigned through a majority vote.

- Aggregation: Like in the case of the language based approach, a different model was trained over each story stem and the decisions made at the level of individual stems were aggregated through WA.

Results

Overall Performance: Figure1 below shows that all models perform significantly better than chance. However, as shown in figure 2, the performance changes from one story stem to the other for individual modalities and their multimodal combination. This confirms the assumption that different stems tend to elicit attachment-relevant behaviours to a different extent.

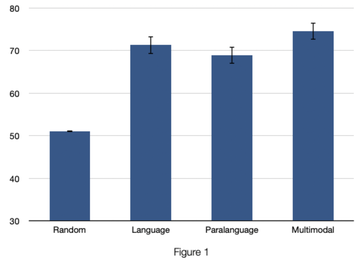


Figure 1

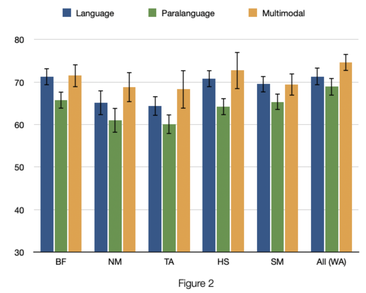
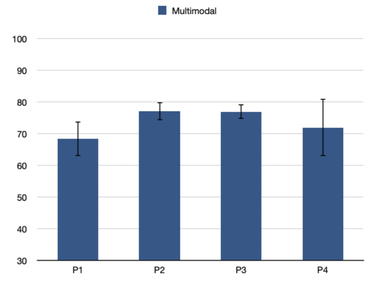


Figure 2

The figure also shows that the aggregation of different stories lead to a significant improvement in the case of the paralanguage thus suggesting that children change the way they tell the stories depending on the stem. However, this is not the case for language as the aggregation does not lead to better performance. Moreover, the presence of two modalities has led to better performance, as shown in figure 2-All(WA), suggesting that paralanguage and language carry complementary information that can help the two modalities to mutually correct each other.

Results

School Level Performance:
The following figure shows how the multimodal approach performs for different age groups.



The difference between level P1 and levels P2 and P3 is statistically significant ($p<0.01$) but the same does not apply to the difference between P1 and P4. While it seems that the performance is increasing when passing from P1 to P2 and P3, this pattern does not seem to remain when passing from level P1 to P4.
One possible explanation 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 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).

Conclusions

- 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%).
- The approach leads to the identification of a large fraction of the insecure children, achieveing a recall of 58.7%.
- The multimodal approach improves over both unimodal approaches, thus showing that the two behavioural channels tend to carry complementary information.
- 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.
- Future work will focus on the inclusion of different modalities, (e.g., facial expressions) and the identification of attachment markers.

Literature cited

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