**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 120 children of age 5-9 years (30 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).

**The Multimodal Approach**

The proposed approach builds upon two modal recognizers: 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 other for paralanguage. The outcomes are combined through Weighted Averaging (WA), i.e., by estimating the probability the unimodal systems attribute to the other for paralanguage. The outcomes are combined through Weighted Averaging (WA), i.e., by estimating the probability the unimodal systems attribute to the other for paralanguage.

**Language-Based Approach**

This approach consists of three main steps: pre-processing, classification, and aggregation.

- **Pre-processing**: Eliminates punctuation, non-alphabetic characters and numbers, stemming, stop-words removal. 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 input sequences into 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.

**Paralanguage-Based Approach**

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

- **Feature Extraction**: The features are extracted using OpenSmile[2] over 33 ms long non-overlapping analysis windows. The feature set consists of 46 basic features and their respective delta regression coefficients, resulting in 32 features that were shown to be effective in emotion recognition. The output of the feature extraction is a sequence of feature vectors \( X = (x_1, x_2, ..., x_T) \). The basic features are as follows:

  - **Root mean square of the energy**
  - **Fundamental frequency**
  - **Mel Frequency Cepstral Coefficients**
  - **Zero Crossing Rate**
  - **Voicing probability**

- **Classification**: A deep network consists of: an embedding layer to convert the input sequences into 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**: The approach leads to the identification of a large fraction of the insecure children.

**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 skills. 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, achieving a recall of 58.7%.
- The multimodal approach improves over both unimodal approaches, thus showing that the two behavioral 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**