ASSESSMENT OF A CHILD'S ENGAGEMENT USING SEQUENCE MODEL BASED FEATURES

Rahul Gupta, Chi-Chun Lee, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), Department of Electrical Engineering, University of Southern California, Los Angeles, CA

ABSTRACT

Detecting and modeling the engagement of a child during an interaction offers meaningful insights into socio-emotional and cognitive state assessment. Previous work has shown that the engagement level of a child during an interaction with a psychologist can be captured from their vocal behavior. In particular global statistical measures on vocal features computed over an entire interaction were associated with the perceived level of engagement. We extend this framework by introducing a new scheme to capture the temporal patterning of vocal features using sequence models of the interacting child-psychologist dyad. We achieve enhanced unweighted accuracies of 73.23% (chance 50.00%) in a classification experiment of distinguishing the most engaged state against others and a three way accuracy of 51.42% (chance 33.33%) in discriminating three levels of perceived engagement using the new set of features.

Index Terms— Behavior signal processing, child engagement, sequence model

1. INTRODUCTION

Several research studies interlink childhood development, speech acquisition, joint attention and engagement [1, 2, 3]. In our previous investigations on child-psychologist interactions (captured in the Rapid-ABC database [4]), we observed that the vocal cues from the child and the psychologist carry useful information in discriminating various levels of perceived child engagement. This was solely based on the global statistics of prosodic, spectral and speech timing features calculated over the entire duration of an interaction. While the prosodic and spectral features carried useful discriminative information, it was found that the speech timing features were the strongest in terms of contributing to the classification accuracy. However since these global measures were calculated at the whole session level, local patterns of vocal behavior during interactions were not completely captured. In the present work, we examine a novel classification scheme that captures and utilizes local variation and patterning of such behavior interaction behavior.

In the proposed model, we use a sequence model of "word units" that are defined on quantized vocal features. Such quantized representation of feature trajectories have been found to be effective in capturing temporal dynamics such as of vocal prosody [5]. The proposed "word units" are used to provide a joint measure on the vocal streams of the child and the psychologist and thus captures the interaction dynamics between the dyad during a session. The model is motivated by the premise that the local patterning of vocal events (e.g., pausing behavior) is informative in assessing the pertinent engagement level beyond behavioral measures of task performance. We combine this model with models based on global statistics of prosodic and spectral features as well as high level features such as total speech length and number of overlaps. The global statistics are calculated on the entire sub-session over which the child's engagement level was assessed by the interacting psychologist. The metrics used to evaluate the performance of such models is unweighted binary accuracy (chance 50%) in a classification task of most engaged state versus others; a finer three way classification task involving three different engagement levels (chance 33.33%) as perceived by the psychologist is also evaluated. We achieve accuracies of 73.23% (relative improvement 2.92%) for the binary case and 51.42% (relative improvement 4.74%) for the three way classification using the new measures. The relative improvements are reported over a model trained only on global statistical measures.

We discuss the database in section 2, and the experimental setup in section 3. Discussion of results and conclusions are presented in section 4 and 5, respectively.

2. DATABASE

2.1. Database description

We use the Rapid-ABC database collected as a part of larger NSF funded study[1] that aims to develop novel computational methods for measuring and analyzing the behavior of children and psychologists during face-to-face social interactions. This database includes a 3-5 minute interaction between a child and a psychologist where the psychologist interacts as well as evaluates the child on five different tasks. These tasks consist of smiling and saying hello, ball play, jointly looking at a book, putting a book on your head as if it is a hat, and smiling and tickling. During these activities the psychologist marks the engagement level of the child by a score of '0', '1' or '2', with '0' indicating the highest engagement level. We use these scores as the outcome variable to be predicted. A recent publication [6] on the same database, gives a more detailed description of the database. [6] uses both the modalities of vision and speech for activity detection and lays the groundwork for linking them to diagnosis and treatment of developmental disorders such as autism.

We use 63 sessions of data from children 9-30 months old for our current study. Our vocal analysis is performed on the audio from a central farfield microphone used in data collection. Since we have 5 tasks per subject, there are $63 \times 5 = 315$ sub-sessions over which we have the engagement scores. 224 of these sub-sessions have a score of '0', 60 a score of '1' and 31 are assigned a score of '2'. Since the child did not need to speak during the interaction, 255 out of the 315 sub-sessions contain child vocal activity. All of these sessions necessarily have psychologist speech.

2.2. Database processing

We first segment the database into speech and non-speech segments using a voice activity detection (VAD) system based on long term

[1] http://www.cbs.gatech.edu/

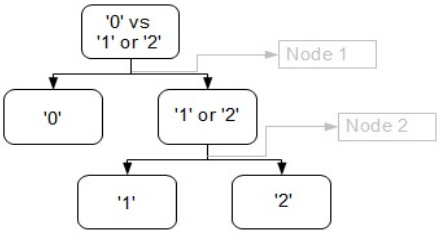

Fig. 1. Classification tree

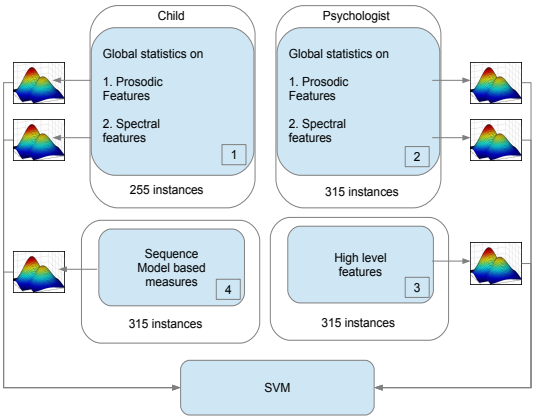

Fig. 2. Classification setup

Table 1. List of features

Feature Category	Source	Features	Functionals (as applicable)
Prosodic Features	Child Psyc.	Pitch, Intensity, Jitter, Shimmer	Mean, Variance, Range, Kurtosis, Quantiles(10%, 25%, 50%, 75%, 90%)
Spectral Features	Child Psyc.	MFCC (13 Bins)	Mean Variance
High level Features	Sub-session length (normalized task-wise) Child speech length, Number of (#) overlaps, Total Speech activity, #Psychologist utterances, #Child utterances		

spectral variability features [7] in every 10 ms window. After obtaining the voiced segments in the database, we perform a semi-supervised speaker segmentation on the dataset. The speech segments for the child are annotated manually using Audacity software [8]. This is followed by the other voiced segments directly attributed to the psychologist, as this database essentially involves interaction between two people.

3. EXPERIMENTS

We train several discriminative models based on the extracted features as discussed below. We use Gaussian mixture models (GMM) as our base classifiers and evaluate the performance using leave-one-session-out (= 5 sub-sessions) cross-validation. In order to determine the parameters on the train set, we perform an inner cross-validation on the train set again by leaving one session out at a time. The classification is performed using a tree structure as shown in figure 1. We first carry out binary classification between states '0' vs '1' or '2' and then between'1' and '2' at node 2 . The entire classification setup is summarized in figure 2. The blocks 1, 2 and 3 use GMM trained on global statistics on various features. We capture the local patterns in the features in block 4. Finally we perform a fusion

Table 2. Accuracies on global statistics of features

Feature Category		Unweighted Accuracy	
		Binary (Tree Node 1)	3-class (Tree Node 1+2)
Prosodic Features	Child	59.01	42.74
	Psyc.	54.30	36.58
Spectral Features	Child	67.15	38.55
	Psyc.	65.17	43.53
High level Features		65.36	45.35

based on stacked generalization [9] using a support vector machine classifier on probability outputs from the individual GMMs.

3.1. Classifiers based on global statistics on features

In the previously proposed classifier scheme [4], we used prosodic, spectral and other high level features from the child and the psychologist and trained a multiple logistic regression model on their statistics over the entire sub-session. In this paper, we change the base classifier to Gaussian mixture models and expand the feature set and perform forward feature selection on the train set to maximize the unweighted accuracy. The features and their functionals are listed in Table 1. The features were extracted using Praat [10] and mean normalized per speaker. The results are presented in Table 2 and they follow a similar trend as observed in [4]. We see that the prosodic features from the child speech are better indicator of engagement as compared to psychologist speech. It is also observed that a few high level features give the best classification accuracy in 3-way classification. This indicates that features of speech duration and the timing patterns of speech from both the speakers also contain useful information with regards to engagement assessment.

We plot the two most informative features for the prosodic and high level feature categories at each of the two nodes in the binary classification tree in figure 3. As is also reflected in the results, most of the prosodic features do not occupy very distinct regions, leading to inter-class confusion. However, the pitch and intensity range for the child and intensity range for the psychologist tend to go higher with increased disengagement as seen in the distribution of node 1 datapoints. A more discriminating pattern is observed in the data point distribution for the high level features, where we see that an increase in the total session length (duration) implies more disengagement for Node 1. Similarly a higher number of child utterances are observed for class '2' as compared to class '1' in the node 2 plot. This suggests that greater amount of speech activity leads to higher disengagement as the tasks require visual joint attention instead of vocal interaction.

3.2. Classifiers based on sequence model (SM) probabilities

As also pointed out in [6], the psychologist structures his behavior as per the child and thus an important part of the engagement process. In this model we intend to capture the patterns in features during dyadic interactions by calculating a measure on each feature stream. We implement a Markov structure on quantized feature levels in this scheme. Thereafter we calculate a metric on feature values based on this Markov structure to capture the local patterns. We define a sequence dictionary on the entire dataset and then classify based on an output probability measure obtained from this model. We explain the sequence dictionary formation and measure calculation based on this dictionary below.

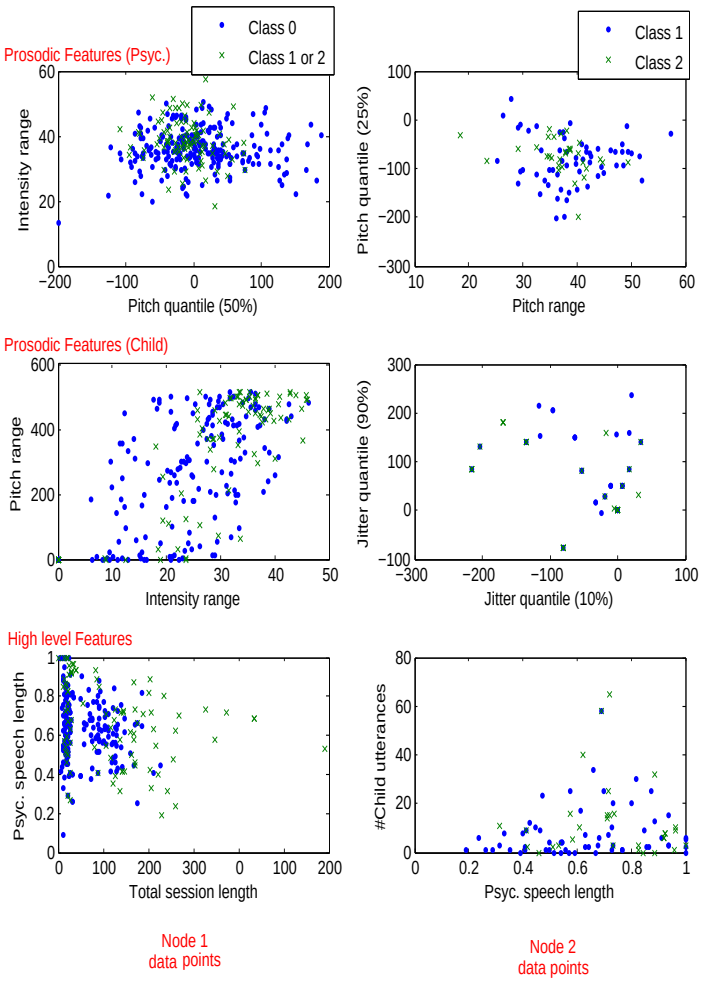

Fig. 3. Classwise plots depicting patterns in features with respect to engagement levels

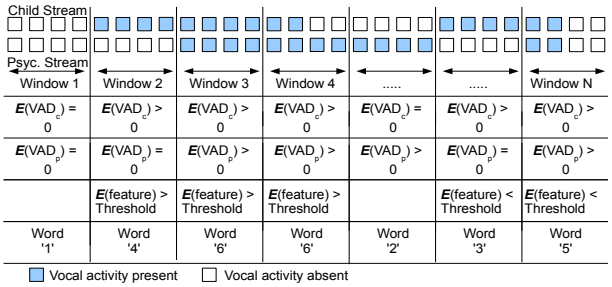

Fig. 4. Dictionary learning to capture local feature patterns

Table 3. Word assignment

Window contains	Assigned word
No vocal activity	1
$\mathrm{E}(\mathrm{VAD}_C) = 0$ & $\mathrm{E}(\mathrm{VAD}_P) > 0$	2
$\mathrm{E}(\mathrm{VAD}_C) > 0$ & $\mathrm{E}(\mathrm{VAD}_P) = 0$) & E(feature) $\leq$ threshold	3
$\mathrm{E}(\mathrm{VAD}_C) > 0$ & $\mathrm{E}(\mathrm{VAD}_P) = 0$) & E(feature) $>$ threshold	4
$\mathrm{E}(\mathrm{VAD}_C) > 0$ & $\mathrm{E}(\mathrm{VAD}_P) > 0$) & E(feature) $\leq$ threshold	5
$\mathrm{E}(\mathrm{VAD}_C) > 0$ & $\mathrm{E}(\mathrm{VAD}_P) > 0$) & E(feature) $>$ threshold	6

Table 4. SM Features and their accuracies

Feautre used in LM	Unweighted Accuracy (%)	
	Binary (Tree Node 1)	3-class (Tree Node 1+2)
Constant	71.51	47.76
Child Pitch	68.25	40.74
Psyc. Pitch	70.38	47.94
Child Intensity	66.60	41.29
Psyc. Intensity	66.60	38.67

3.2.1. *Sequence dictionary creation*

Initially, we segment the session into small non-overlapping segments on the VAD output of window length W. Each of these segments is assigned a "word" based on the outputs from the VAD, speaker segmentation and a pre-selected feature on one of the speakers. VAD_C and VAD_P indicate the presence or absence of child and psychologist vocal activity, respectively. The word assignment strategy for a feature defined on child speech is listed in Table 3. An example of a sample session with $W = 4$ and a feature chosen on child voice is shown in figure 4. The threshold in the figure is used to quantize the window-wise mean of the features into two levels and is chosen as mean of the feature over the entire sub-session. Using this thresholding scheme, we aim to capture the inherent distribution of these features during the interaction. Quantizing the features into two bins can be viewed as approximating them by a Bernoulli distribution with one outcome representing the window-wise mean to lie above a threshold and the other below. E is the sample expectation operator over the pertinent window and N is the total number of windows in the session. After obtaining such a sequence of "words" on the training data, we develop an n-gram sequence model (SM). An n-gram models the probability of sequence of words, similar to language modeling. The SM is trained based on the frequency count for each of the n-gram sequences ($\#(w_k^{S_t}/w_{k-1}^{S_t}, .., w_{k-n+1}^{S_t})$) for all windows $S_t \in$ Training set (equation 1). These word counts are normalized by the total number of words from all the training sessions. N_t is the number of windows in the t^{th} session.

$$\mathrm{SM}(w_k/w_{k-1}, .., w_{k-n+1}) = \frac{\sum_{t \in \text{Train Set}} \#(w_k^{S_t}/w_{k-1}^{S_t}, .., w_{k-n+1}^{S_t})}{\sum_{t \in \text{Train Set}} N_t} \quad (1)$$

3.2.2. *Measure generation from the sequence dictionary*

We generate a set of measures from the above SM trained on the entire train set. The measure for a sample session S is defined as the n-gram count in that session multiplied by the SM probabilities (equation 2). This gives the relative occurrence of each n-gram pattern in a session. These measures represent the likelihood of an n-gram sequence as is generated from a particular session. We hypothesize that the distribution of chosen features and the interaction pattern would be different for different levels of engagement. In such a case, these differences will be reflected in the likelihoods for the n-gram sequences.

$$\mathrm{M}(w_k^S/w_{k-1}^S, .., w_{k-n+1}^S) = \mathrm{SM}(w_k/w_{k-1}, .., w_{k-n+1}) \times \#(w_k^S/w_{k-1}^S, .., w_{k-n+1}^S) \quad (2)$$

For example, an n-gram trained on 6 such words leads to 6^n such measures. We train a discriminative Gaussian mixture model on them. We use a bi-gram model for the feature generation and did not define dictionaries with multiple features as more complex n-gram models led to sparse SM, which decreased the accuracies in our experiments. The features and the corresponding results are

Table 5. Classification accuracies after fusion

Feature Source	Selected Model	Class-wise Accuracy				Unweighted Accuracy (2 class)	Unweighted Accuracy (3 class)
		Class '0'	Class ∼'0' (Node 1)	Class '1'	Class '2'		
Global statistics on features	High Level features[1,2]	78.57%	63.73%			71.15%	
		78.57%		30.00%	38.70%		49.09%
Feature based on SM	Constant[1,2], Child Pitch[1,2] Psyc. Pitch[1,2], Child Intensity[2]	77.23%	69.23%			73.23%	
		77.23%		33.33%	32.25%		47.60%
Fused Model	Constant[1,2], Child Pitch[1,2] Psyc. Pitch[1,2],Child Intensity[1,2], High level features[2]	77.23%	69.23%			73.23%	
		79.01%		33.33%	41.93%		51.42%

listed in Table 4. We also train a model without using any feature to include the pure vocal interaction pattern between the two speakers. This reduces the number of words to assignments to 4 from 6 as in Table 3. This is same as setting the feature to be a constant and thus is referred to as "constant" feature in Table 4. We set $W = 10$ in the experiments.

From the results in Table 4, we observe that we can obtain a good classification accuracy by incorporating feature dynamics along with the interaction pattern between the speakers. A plot of occurrence count of "word 4" given "word 4" and "word 4" given "word 5" is shown in figure 5. It can be seen that pitch and intensity tend to have higher values for both psychologist and child for the more disengaged classes.

3.3. Feature fusion

We obtain our final label based on a support vector machine (SVM) classifier trained on the posterior probabilities obtained from the Gaussian mixture models trained at node 1 and 2. During the training of such a SVM, we set the cost of misclassification for each of the classes inversely proportional to the number of class instances due to their uneven distribution. We initially observe the fusion results on the global and SM models individually and then fuse all the models. We perform a brute-force search as to what discriminative models should be finally used to train the SVM model for optimal binary and 3-class accuracies. The selected models and the corresponding accuracies are listed in Table 5.

4. DISCUSSION

We observe that for the global model, the high level features by themselves are the most informative. The features based on the SM probabilities are stronger for binary classification with a higher accuracy for '∼0' class, but worse in the 3-way classification. The best binary classification uses the SM features only. This shows that the use of SM measures is better able to make the broad distinction of being engaged or not, which is otherwise diluted in the calculation of the global features. Note that in the 3-way classification the class '0' accuracy is higher as compared to other classes due to the nature of the classification. The SVM classifier first balances accuracies between class '0' and '∼ 0'. Then it classifies class '1' from '2'. During the final fusion from all the features, all the features selected in individual models help and we observe the best accuracy for all the classes. This suggests that even though the SM measures have less discriminative power in classifying between the two higher disengagement states, they carry complementary information to the global features in capturing the characteristics depicted in class '0' and class '2' sub-sessions particularly. This indicates that the engagement level perceived by the psychologist depends on not only the outcome of the overall perception of the psychologist, but also how local events unfold during the interaction.

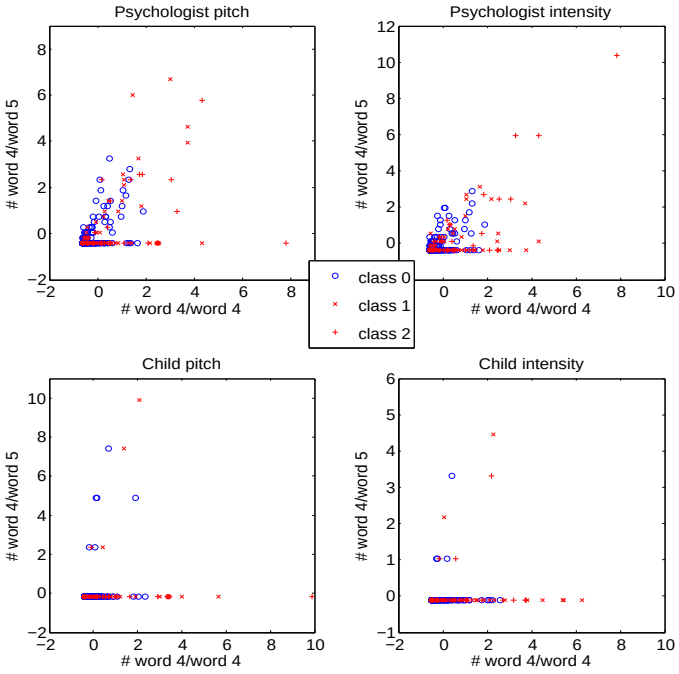

Fig. 5. Bigram counts of word occurrence for the three levels of perceived engagement

5. CONCLUSION

In this work, we examined a novel sequence modeling scheme to assess the engagement level of children during dynamic interactions. We observe that not only do global measures calculated over the entire session carry discriminative information about the engagement level as perceived by the psychologist, the local dynamics of feature patterns is also informative. We further show that we can fuse the two feature sources to achieve our best model.

As future work, one can look into better quantization scheme for such sequence dictionary formation. As of now, we chose the mean of global features as the threshold. However this is not necessarily optimal for classification. Apart from this, more complex sequence models with higher order n-grams and multiple features can be tested with larger databases that are not limited by data sparsity. Finally, one can also apply better smoothing techniques that are prevalent in language modeling before feature extraction to further smooth the sequence model.

[1]optimal for node 1 [2] optimal for node 2

6. REFERENCES

[1] D. McCarthy, "Language development.," *Clark University Press*, 1931.

[2] K.A. Loveland and S.H. Landry, "Joint attention and language in autism and developmental language delay," *Journal of autism and developmental disorders*, vol. 16, no. 3, pp. 335–349, 1986.

[3] J. Piaget, "Part i: Cognitive development in children: Piaget development and learning," *Journal of research in science teaching*, vol. 2, no. 3, pp. 176–186, 1964.

[4] R. Gupta, C.C. Lee, D. Bone, Rozga A., S. Lee, and S. Narayanan, "Acoustical analysis of engagement behavior in children," *Workshop on Child, Computer and Interaction, Portland, Oregon*, 2012.

[5] V.K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech & Language*, vol. 23, no. 4, pp. 407–422, 2009.

[6] J. Rehg et al., "Decoding children's social behavior," in *Computer Vision and Pattern Recognition, 2013. CVPR. IEEE Conference on*.

[7] P.K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 600–613, 2011.

[8] D. Mazzoni and R. Dannenberg, "Audacity [software]. pittsburg," 2000.

[9] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[10] W. Boersma and D. Weenink, "Praat software," *Amsterdam: University*, 2006.

[11] M.P. Black, P. Georgiou, A. Katsamanis, B. Baucom, and S. Narayanan, "You made me do it:classification of blame in married couples interactions by fusing automatically derived speech and language information," *Proc. Interspeech*, 2011.