Proceedings of Workshop on
Child, Computer and Interaction
(WOCCI 2012)

Acoustical analysis of engagement behavior in children

Rahul Gupta[1], *Chi-Chun Lee*[1], *Daniel Bone*[1], *Agata Rozga*[2], *Sungbok Lee*[1], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, Viterbi School of Engineering,
University of Southern California, Los Angeles, California, USA

[2]School of Interactive Computing,
Georgia Institute of Technology,
Atlanta, Georgia, USA

guptarh@usc.edu, chiclee@usc.edu, dbone@usc.edu, agata@gatech.edu,
sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

In this work we analyze the expressive manifestation of a child's engagement behavior on his speech as well as in the speech of psychologist interacting with the child. Visual cues such as facial gestures and gaze are known to be informative of engagement, but here, we examine the less studied speech cues of the children's non-verbal vocalizations. We study the spectral, prosodic and duration features obtained from the child and the psychologist's vocal data. We observe that these measures carry discriminative power in assessing specific engagement levels of the children (49.2% accuracy in classifying 3 levels of engagement compared to 33% chance accuracy). We also present our results as a detection task for disengagement with precision, recall and f-measure of .70, .42, .53, respectively. The unweighted accuracy for binary classification between engagement and disengagement is 62.9%. Our results suggest that vocal cues bear useful information in capturing the state of engagement in speech, indicating that speech can play an effective role in engagement assessment.

Index Terms: Child engagement, Rapid-ABC, autism

1. Introduction

Research studies on childhood development are broad and multifaceted and have focused on several aspects of motor, speech, language and socio-emotional development [1, 2, 3, 4, 5]. Considerable work has also focused on the interplay between these developmental aspects. For instance, the study of joint attention has been associated with speech and language development as well as social cognition [6, 7]. Effects of joint attention on spontaneous speech for children has revealed important insights (e.g. increase in spontaneous speech by teaching joint attention skills [8, 9]). Particularly for disorders like autism, patterns of this relationship are especially of interest when behavior is deemed atypical [3, 10, 11, 12, 13, 14]. Engagement can be closely related to joint attention as it is the process of sharing ones attention and interest with another person using gestures or gaze [15, 16], (for example in child play and interaction with the parents [17, 18]). Our work focuses on engagement which can help inform the joint attentional aspects of social interactions with a children.

For our experiments, we hypothesize that the speech produced in the process, including the speech of the clinician [19], also contains information helpful in determining one's level of engagement. For this purpose, we use the Rapid-ABC database, which comprises short interactions between a child and a clinician involved in a series of scripted activities. These activities elicit and study the social communicative behavior of the children.

We explore the role of vocal cues, including that of the interlocutor, in assessing engagement. Our experiments exclusively focus on vocal cues (speech) and its relationship with the engagement level of children. Most of the vocalizations produced by the children in the database are non-verbal as they are 9-30 months old. The clinician interacts with the child, gauging his performance on a defined set of tasks as well as rating the effort required to engage the child as '0','1' or '2', with '0' indicating little effort. We fuse the information from acoustic-prosodic and spectral features from the child and the clinician speech, and speech duration features to obtain a final engagement score for the child. For this three-class classification, we achieve an unweighted accuracy of 49.2% over chance accuracy of 33.3%. We also pose the question of classifying engagement from speech as "disengagement" detector where the classes '1' and '2' are representative of disengagement, achieving an unweighted accuracy of 62.9%. Thus our experiments explore the role of speech in defining engagement levels in children using engineering models, which when fused with cues from other sensory channels (e.g. visual) may provide additional information to the psychologist.

In section 2 we describe the research methodology. Section 3 contains the experimental setup, results, discussion followed by conclusions in section 4.

2. Research Methodology

2.1. Database

We use a set of data collected as part of a larger NSF funded study that aims to develop novel computational methods for measuring and analyzing the behavior of children and adults during face-to-face social interactions. In particular, we report on behavioral data collected as part of the Rapid-ABC, a 3-5 minute interactive assessment designed to elicit key social communicative behaviors, including social attention, back-and-forth interaction and nonverbal communication. The database consists of mostly non-verbal children, 9-30 months old. The initial phase of the database is comprised of 52 subjects and 54 sessions (sessions repeated for 2 children). Five different tasks designed to elicit expected responses are conducted in each session. These tasks are smiling and saying ”hello”, ball play, jointly looking at a book, putting on a book on your head as if it is a hat, and smiling and tickling.

There are two specific guidelines to score the child in each of these five activities. One set of scores mark certain actions taken by the child during the session. For instance, a child is given a ‘+’ or ‘-’ depending on if s/he makes eye contact and smiles or not. For each subtask, the psychologist also notes whether the child was easy to engage by a mark of ‘0’,‘1’ or ‘2’. A score of ‘0’ is given if the interaction required no to little effort by the psychologist and the child was ready and eager to be engaged. ‘1’ represents some effort on the part of the psychologist due to the child’s shyness or distractability. ‘2’ is given for extensive effort by the psychologist or if the child is very fussy and refuses to interact. We use these scores for engagement and they are defined as class ‘0’, ‘1’ or ‘2’ in our classification experiment.

The database has several modalities including video, audio and electro-dermal activity recording. We use audio from lapel microphones for our analysis. We examine 50 children in separate sessions ($\times$5 subtasks = 250 sub-sessions) from the database because the lapel mike channels were corrupted for two of the children. 171 of these sub-sessions are marked ‘0’ on the engagement scale, 49 as ‘1’ and 30 as ‘2’ (Table 2). Hence, we observe that the instances of the child being disengaged are rare as compared to the instances where the child was given a score of ‘0’. Table 1 shows the transition matrix for engagement scores between consecutive sub-sessions in a session, as derived from the database. Class ‘0’ can be seen as a ground state for engagement because it is the most likely transition given any current state. This fact motivates our experiment on classifying the most engaged state from the other two states.

Table 1: *Transition matrix for engagement levels*

		Next Class		
		Class ‘0’	Class ‘1’	Class ‘2’
Current Class	Class ‘0’	.75	.17	.08
	Class ‘1’	.61	.24	.15
	Class ‘2’	.50	.14	.36

2.2. Database Annotation

The first author marked the start and end time of the child speech using Audacity software [20]. As the tasks don’t require the child to respond vocally, he may be silent throughout the entire sub-session. We find 194 of the 250 sub-sessions contain some child’s vocal activity. A breakdown of child engagement scores with the occurrence of child speech is shown in Table 2.

Table 2: *Engagement level vs Presence of child speech.*

Sessions	**Class ‘0’**	**Class ‘1’**	**Class ‘2’**
All Sessions	171	49	30
With child speech	125	42	27
Without child speech	46	7	3

From the table, we observe that class ‘1’ and class ‘2’ are more likely to occur in the presence of child speech. In the absence of child speech, there are 10 instances of class ‘1’ and class ‘2’ combined together as compared to 46 instances from class 0 (odds ratio 4.6). We have 69 and 125 such instances in the presence of child speech (odds ratio 1.81). We use this intuition when we determine our feature set in the form of speech duration features.

3. Experimental Setup

In this section we explain the classification approach to identify engagement levels using speech from the child and the psychologist. We break up our analysis into three parts in terms of acoustic features. Such feature extraction and the classifier setup is explained section-wise below.

3.1. Feature Extraction

We extract acoustic-prosodic, spectral and speech duration features (Table 3) for both child and the psychologist, using Praat software [21]. The spectral and the prosodic features are mean normalized for each speaker. Then, means and the variances are calculated over all the utterances of each speaker in each sub-session. Since we do not have manual annotation for the psychologist

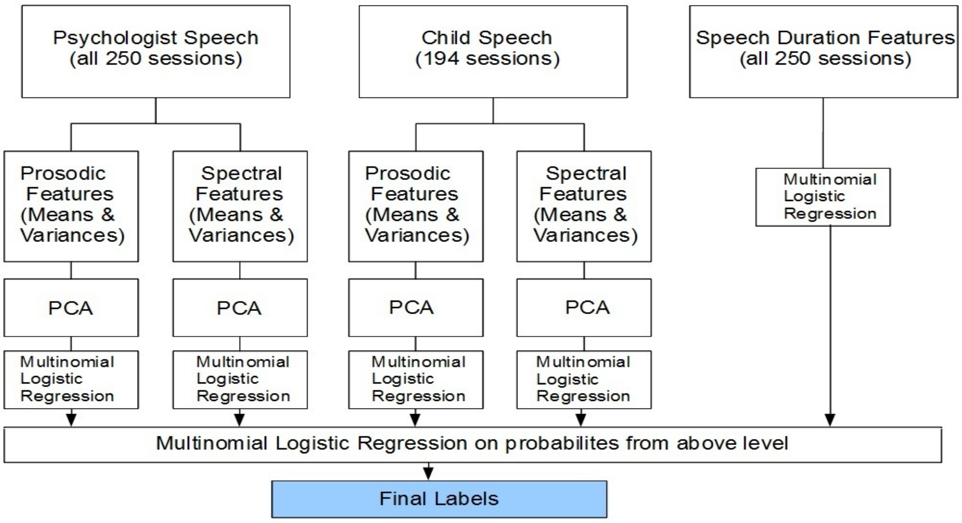
Psychologist Speech
(all 250 sessions)
Child Speech
(194 sessions)
Speech Duration Features
(all 250 sessions)
Prosodic Features (Means & Variances)
Spectral Features (Means & Variances)
Prosodic Features (Means & Variances)
Spectral Features (Means & Variances)
Multinomial Logistic Regression
PCA
PCA
PCA
PCA
Multinomial Logistic Regression
Multinomial Logistic Regression
Multinomial Logistic Regression
Multinomial Logistic Regression
Multinomial Logistic Regression on probabilites from above level
Final Labels

Figure 1: *Classification setup*

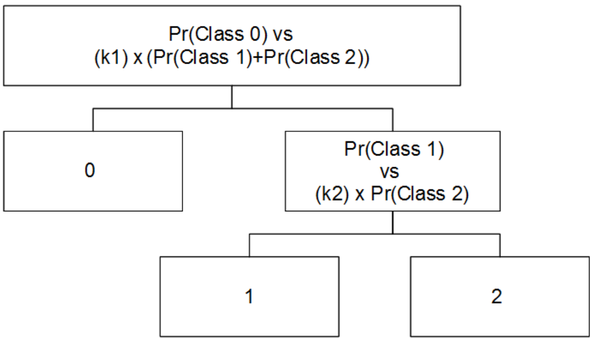

Figure 2: *Class assignment tree*

Table 3: *Feature description.*

	Feature	Statistics used
Spectral Features	MFCC	μ, σ
Prosodic Features	Intensity	μ, σ
	Pitch	μ, σ
	Jitter	μ, σ
	Shimmer	μ, σ
Speech Duration Features	Child Speech Duration	As % of session duration
	Psychologist Speech Duration	As % of session duration

speech, we design an energy-based voice activity detector. The audio from the psychologist lapel microphone is low pass filtered and compared to a threshold to detect psychologist speech. All the instances of voice activity overlapping with child speech are removed so that they are not marked as psychologist speech. By the virtue of this, speech overlaps may appear as just the child speech. Also, the fraction of each sub-session containing child and psychologist speech (as two different features) are used to capture our hypothesis of a difference in engagement level based on the amount the child speaks.

3.2. Classification setup

We perform our experiments in a cross-validation using 8:1:1 ratio to determine the train, development and test set over the 50 sessions. We train separate multinomial logistic regression models for child, psychologist and speech duration features, after performing dimensionality reduction on the feature set using principal component analysis. A breakdown of the classification framework is given in Figure 1 and the setup is explained below.

3.2.1. Child Speech

We train multinomial logistic regression on 194 sessions where the child speech occurs. For cross-validation, we leave all the sub-sessions in 5 sessions for testing and sub-sessions from other 5 as development set. The other 40 sessions are used for training. Models are trained separately on prosodic features (dimensionality 8) and spectral features (dimensionality 26) after reducing the dimensionality using principal component analysis. PCA is particularly important for spectral features as we have fewer data samples, given the dimensionality of 26. The number of principal components used is tuned on the development set. We use unweighted accuracy as our metric because of the heavy bias in the database. Unweighted accuracy is reported for each feature set individually as well as after fusion of the two feature sets. For fusion,

Table 4: Classification Accuracies for each feature channel

Feature Source	**Feature Set**	**Unweighted Accuracy**	Effective for
Child	Chance	33.3%	
	Spectra	34.1%	Class ‘0’
	Prosody	32.6%	Class ‘1’
	Fused model	43.6%	
Psychologist	Chance	33.3%	
	Spectral	36.3%	Class ‘0’
	Prosody	36.4%	Class ‘0’
	Fused model	37.0%	
Speech duration		41.9%	Class ‘0’,‘1’

we multiply weighted probabilities from the two models, with weights tuned on the development set. For final assignment of classes, we break down the process into two steps. The first classification is class ‘0’ vs ∼‘0’. We directly compare the probabilities from class ‘0’ and the sum of probabilities from class ‘1’ and ‘2’ weighted by a factor k1 (Figure 2). In the next step, if the class is obtained to be ∼‘0’, we perform a probability comparison between class ‘1’ and ‘2’, weighting the probabilities by k2. The parameters k1 and k2 are tuned on the development set.

3.2.2. Psychologist Speech

Since the psychologist speaks in all the sessions, we have his/her speech available in all the 250 sub-sessions. We perform a similar split and similar classification as above and the results are reported for all the 250 sub-sessions. These features become critical in the absence of child speech.

3.2.3. Speech duration features

Features in this case are the speech durations of the child and the psychologist expressed as a percentage of sub-session time length. The experimental setup is again same as above and the child speech time is set to 0 for the sessions where it is absent.

The final fusion is done in a stacked generalization framework [22]. We take the probabilities from each of the five systems (2 spectral + 2 prosody + 1 time duration) and train a regression model on the development set. The probabilities from the test set are then used to determine the final label. We assign the final label as shown in Figure 2. Since the data is biased, we multiple the probability thresholds by k1 and k2, tuned on the development set for maximum unweighted accuracy.

3.3. Results

We report the results for classification on child speech, psychologist speech and the interaction features in Table 4. Note that the results for child speech are not directly comparable with psychologist and interaction features because of the fewer number of sub-sessions. The final results after fusion are presented in Table 5. We also report the results for the binary classification task for disengagement where class ‘1’ and ‘2’ are merged to represent disengagement.

3.4. Discussion

We discuss the results for each of the classifiers below.

3.4.1. Child features

For the child features, we observe that whereas the spectral features do well for class ‘0’, the prosody features perform well for class ‘1’. Even though each of these features perform close to chance for the unweighted accuracy, they carry complementary information as to the class accuracies. Hence we see an overall boost after fusion.

3.4.2. Psychologist features

The psychologist features are very effective for class ‘0’ for all the cases. After fusion, class ‘0’ is classified almost perfectly, whereas there is some power in class ‘1’. We do not observe any power in classification for class ‘2’. Poor performance of these features can also be attributed to the energy based VAD. Since this form of VAD is not very robust, we are likely to include other sources of noise at times and might even exclude some of the psychologist speech.

3.4.3. Speech duration features

These features are also most discriminative for class ‘0’, but also perform well for class ‘1’. The discriminative power of these features indicates that simple features like these can help us in analyzing the engagement behavior of children. This is also in agreement with our previous intuition of using speech length as features. From all above feature sets, we observe that the classification is mostly biased towards class ‘0’. Class ‘2’ is weakest in classification. To address the problem of bias in the accuracy, we add class priors in our final fusion model.

3.4.4. Fusion framework

The fusion framework is tuned to maximize the unweighted accuracy. We observe that after fusion, the class accuracies are more or less balanced. This suggests that

Table 5: Classification Accuracies after fusion from the three feature channels

Feature Source	**Class-wise Accuracy**				**Unweighted Accuracy** (3 class)	**Unweighted Accuracy** (2 class)
	Class ‘0’	Class ∼‘0’ (First node in Figure 2)	Class ‘1’	Class ‘2’		
Fused model	56.1%	69.6%				62.9%
			44.9%	46.7%	49.2%	

Table 6: Classification Accuracies after fusion from the three feature channels

Feature Source	**Class-wise Accuracy**		**Unweighted Accuracy** (2 Class)	**Unweighted Accuracy** (3 Class)
Class ‘0’	Class ∼‘0’			
56.1%	69.6%		62.9%	
	Class ‘1’	Class ‘2’		
	44.9%	46.7%		49.2%

the feature set, after fusion carries fairly equal discriminative power for all the classes. From the confusion matrix (Table 6), we observe that class ‘1’ is mostly confused with class ‘0’ and class ‘2’ with class ‘1’. This suggests that stronger classification between class ‘0’ and the other two classes can help improve the accuracy. Also within the class ‘1’ and ‘2’, better algorithms and features need to be devised to capture the differences between the two levels of disengagement.

Table 7: *Confusion Matrix with class weighting*

		True Class		
		Class ‘0’	Class ‘1’	Class ‘2’
Predicted Class	Class ‘0’	96	17	7
	Class ‘1’	48	22	9
	Class ‘2’	27	10	14
Total		171	49	30

3.4.5. *Disengagement detection*

We can pose this problem as disengagement detection task since class ‘1’ and ‘2’ are rarer in comparison to ‘0’ and class ‘0’ acts as the ground state (Table 1). We represent the Table 6 in this view, where we merge the class ‘1’ and ‘2’ (Table 7). The precision, recall and f-measure for such a detection problem are .70, .42 and .53, respectively. Also, the unweighted accuracy for such a binary classification task is 62.9%. This suggests that even though we have three discrete classes for the engagement levels, there is a minute difference between the class ‘1’ and class ‘2’ as compared to class ‘0’. Since the assignment of these classes is based on subjective judgment, it becomes challenging to draw defined boundaries between these two classes.

Table 8: *Confusion matrix with class weighting after merging class ‘1’ and ‘2’*

		True Class	
		Class ‘0’	Class ‘1’ and ‘2’
Predicted Class	Class ‘0’	96	24
	Class ‘1’ and ‘2’	75	55

From the results, we observe that though each of the feature channels are weak in classification, we get a boost after making a fused decision. Even though these accuracies are not very high in magnitude, they perform better than chance all the time, indicating that the selected features are capable of providing knowledge in analyzing the phenomena of engagement. However, the setup of the current experiment can be refined further to extract more information from these features. We can use a better voice activity detection to segment the psychologist speech more robustly. Additionally, the classification framework can be tuned further to improve the accuracy in the perspective of disengagement detection.

4. Conclusion

We show that vocal cues are informative of a child’s engagement, even though engagement is highly reflected in visual behavior. We process the speech from a dyadic interaction between the child and psychologist in Rapid-ABC sessions and predict the engagement scores as given by the psychologist. Since the psychologist plays the role of both interactor and evaluator, we expect this to reflect in his speech in addition to the child speech, whose engagement is being evaluated. We, through our experiments, show that the selected vocal cues do contain some expressive power and this goes up with the information fused. Hence, in this paper we make an effort to link the vocal cues in children still in the phase of speech development to their engagement levels, which further opens up more exploratory windows in analyzing social behavior of children during interactions.

As a next step, we can look into the interaction dynamics between the psychologist and the child. In the stated experiments, we treat these two sources separately and perform fusion without studying their interrelation. In the future, other turn taking features as overlap of speech, occurrence of pauses can be studied. We can also observe the child session as a whole and analyze the temporal evolution of engagement over the sessions, rather than studying them independently. As the data is multi-modal, we can study the other features from video and electro-dermal activity and observe their correlation with speech as well as engagement. As the psychologist is acting as both the interactor and evaluator, this demands a closer scrutiny on his behavior and better cues to predict the child engagement. Finally the temporal evolution of engagement and detection of salient events in speech as well as other modalities within the sub-sessions themselves can be studied to achieve further insights.

5. References

[1] D. McCarthy, "Language development in children." 1946.

[2] M. Shatz and R. Gelman, "The development of communication skills: Modifications in the speech of young children as a function of listener," *Monographs of the society for research in child development*, pp. 1–38, 1973.

[3] K. Loveland and S. 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.

[4] M. Ramscar and N. Gitcho, "Developmental change and the nature of learning in childhood," *Trends in cognitive sciences*, vol. 11, no. 7, pp. 274–279, 2007.

[5] 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.

[6] C. Moore and P. Dunham, *Joint attention: Its origins and role in development*. Lawrence Erlbaum, 1995.

[7] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," *Industrial Electronics, IEEE Transactions on*, vol. 50, no. 4, pp. 636–643, 2003.

[8] P. Mundy, M. Sigman, and C. Kasari, "A longitudinal study of joint attention and language development in autistic children," *Journal of Autism and developmental Disorders*, vol. 20, no. 1, pp. 115–128, 1990.

[9] D. Baldwin, "Understanding the link between joint attention and language," *Joint attention: Its origins and role in development*, pp. 131–158, 1995.

[10] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, "Behavioral manifestations of autism in the first year of life," *International Journal of Developmental Neuroscience*, vol. 23, no. 2, pp. 143–152, 2005.

[11] A. Wetherby, N. Watt, L. Morgan, and S. Shumway, "Social communication profiles of children with autism spectrum disorders late in the second year of life," *Journal of Autism and Developmental Disorders*, vol. 37, no. 5, pp. 960–975, 2007.

[12] S. Ozonoff, A. Iosif, F. Baguio, I. Cook, M. Hill, T. Hutman, S. Rogers, A. Rozga, S. Sangha, M. Sigman *et al.*, "A prospective study of the emergence of early behavioral signs of autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 49, no. 3, pp. 256–266, 2010.

[13] C. Whalen, L. Schreibman, and B. Ingersoll, "The collateral effects of joint attention training on social initiations, positive affect, imitation, and spontaneous speech for young children with autism," *Journal of autism and developmental disorders*, vol. 36, no. 5, pp. 655–664, 2006.

[14] F. Levy, M. McLaughlin, C. Wood, D. Hay, and I. Waldman, "Twin-sibling differences in parental reports of adhd, speech, reading and behaviour problems," *Journal of Child Psychology and Psychiatry*, vol. 37, no. 5, pp. 569–578, 1996.

[15] N. Emery, E. Lorincz, D. Perrett, M. Oram, and C. Baker, "Gaze following and joint attention in rhesus monkeys (macaca mulatta)." *Journal of Comparative Psychology; Journal of Comparative Psychology*, vol. 111, no. 3, p. 286, 1997.

[16] S. Langton, R. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in cognitive sciences*, vol. 4, no. 2, pp. 50–59, 2000.

[17] S. Landry and M. Chapieski, "Joint attention and infant toy exploration: Effects of down syndrome and prematurity," *Child Development*, pp. 103–118, 1989.

[18] C. Whalen and L. Schreibman, “Joint attention training for children with autism using behavior modification procedures,” *Journal of Child Psychology and Psychiatry*, vol. 44, no. 3, pp. 456–468, 2003.

[19] D. Bone, M. Black, C. Lee, M. Williams, P. Levitt, S. Lee, and S. Narayanan, “Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] D. Mazzoni and R. Dannenberg, “Audacity [software],” 2000.

[21] W. Boersma and D. Weenink, “Praat software,” *Amsterdam: University*, 2006.

[22] D. Wolpert, “Stacked generalization*,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.