Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 47–66

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Analysis of engagement behavior in children during dyadic interactions using prosodic cues☆

Rahul Gupta*, Daniel Bone, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Avenue, Los Angeles, CA 90089, USA

Received 10 October 2014; received in revised form 26 July 2015; accepted 16 September 2015
Available online 23 October 2015

Abstract

Child engagement is defined as the interaction of a child with his/her environment in a contextually appropriate manner. Engagement behavior in children is linked to socio-emotional and cognitive state assessment with enhanced engagement identified with improved skills. A vast majority of studies however rely solely, and often implicitly, on subjective perceptual measures of engagement. Access to automatic quantification could assist researchers/clinicians to objectively interpret engagement with respect to a target behavior or condition, and furthermore inform mechanisms for improving engagement in various settings. In this paper, we present an engagement prediction system based exclusively on vocal cues observed during structured interaction between a child and a psychologist involving several tasks. Specifically, we derive prosodic cues that capture engagement levels across the various tasks. Our experiments suggest that a child's engagement is reflected not only in the vocalizations, but also in the speech of the interacting psychologist. Moreover, we show that prosodic cues are informative of the engagement phenomena not only as characterized over the entire task (i.e., global cues), but also in short term patterns (i.e., local cues). We perform a classification experiment assigning the engagement of a child into three discrete levels achieving an unweighted average recall of 55.8% (chance is 33.3%). While the systems using global cues and local level cues are each statistically significant in predicting engagement, we obtain the best results after fusing these two components. We perform further analysis of the cues at local and global levels to achieve insights linking specific prosodic patterns to the engagement phenomenon. We observe that while the performance of our model varies with task setting and interacting psychologist, there exist universal prosodic patterns reflective of engagement.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Engagement; Prosody; Global level cues; Local level cues; Classifier decision fusion

1. Introduction

During childhood an individual develops critical social, physical, psychological and cognitive skills and abilities. This development is affected by several factors including society (Walker et al., 2007; Davie et al., 1972), family (Biller, 1993; Egeland and Farber, 1984) and peers (Dodge et al., 2003). Furthermore, developmental changes are reflected in behavioral aspects such as joint attention (Akhtar et al., 1991; Tomasello and Farrar, 1986), and ability to

☆ This paper has been recommended for acceptance by Martin Russell.

* Corresponding author. Tel.: +1 012135871103.
E-mail address: guptarah@usc.edu (R. Gupta).

http://dx.doi.org/10.1016/j.csl.2015.09.003
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

engage, among others (Cloward et al., 1960; Göncü, 1999; Morrissey-Kane and Prinz, 1999). Quantitative assessment of these behavioral aspects, while very challenging, can provide tools for understanding important aspects of child development, both typical and atypical. In turn, this can further inform intervention methods targeted toward assisting healthy child development. Several approaches for quantifying a child's behavioral aspects such as social and language development (Volkmar et al., 1993; Coplan and Gleason, 1990) exist and we note that their specific type and nature is very dependent on, and tailored to the corresponding behavior of interest. Given the vast heterogeneity and variability in developmental trajectories, especially in the presence of neuro-cognitive and behavioral disorders, there is an imminent need for quantitative methods and analysis tools that can help shed further light into developmental behavioral processes and mechanisms.

Understanding and quantitatively characterizing the engagement patterns of a child, a core behavioral construct, can be useful for both diagnostics and intervention design. Child engagement is defined as the child being involved with his/her environment in a contextually appropriate manner (McWilliam and Casey, 2008). Engagement is a complex internal state externalized and reflected in several modalities including face and body language (Xu et al., 2010; Sanghvi et al., 2011), speech (Yu et al., 2004; Manning et al., 1994) and physiology (Nes et al., 2005). Several studies suggest that a greater engagement has a constructive impact on a child's development (McWilliam et al., 2003; Taylor et al., 2003). For instance, investigations by de Kruif and McWilliam (1999) suggest positive multivariate relationships between developmental age and observed child engagement, where the developmental age is determined over personal–social, adaptive, communication, motor and cognitive domains (Newborg et al., 1984). Göncü (1999) has reported on the impact of a child's engagement during social activities to his/her development and underscores the importance of an interdisciplinary approach to such an endeavor. The importance of engagement is also emphasized in the study of children with developmental disorders like autism (Delano and Snell, 2006; Poulsen and Ziviani, 2004) aiming to enhance behavioral intervention methods (Kasari et al., 2010; Rogers, 2000). Furthermore, improved engagement is associated with success of organizations like child care centers (Maher Ridley et al., 2000) and schools (Skinner and Belmont, 1993). Methods of intervention exist to improve child engagement in different settings such as school (Skinner et al., 1990), parent–child interactions (Casey and McWilliam, 2005) and play with peers (Cielinski et al., 1995).

Given that child development and engagement are strongly coupled, several schemes have been proposed to measure engagement. Yatchmenoff (2005) proposes to quantify engagement in child protective services by categorizing it into five dimensions of receptivity, expectancy, investment, mistrust and working relationship. Kishida and Kemp (2006) have proposed measures of engagement specific to practitioners as opposed to researchers motivated by their relation to practicality, sensitivity to the participants and ability to measure across the span of activity types. Libbey (2004) measured engagement of children in schools by defining a few school connectedness measures based on conceptual interrelatedness. Other studies such as in Read et al. (2002) view engagement as a dimension of a higher order construct and propose measures of engagement as a subcomponent to analyze the construct. However, these studies do not extend to the cases involving natural interaction with children and are often limited to artificial settings. Moreover, given that engagement is a latent internal state inferable only using observed cues, a majority of studies rely on a subjective measurement of engagement. Such measures are susceptible to several uncertainties introduced by variability in interaction settings, interpersonal differences, inconsistencies across subjective judgments and even the operational definition of engagement.

We aim to address the need for an objective engagement behavior quantification method that is robust to the variations in environmental parameters. We perform a study in which children interact with a psychologist while performing different socio-cognitive tasks. The child–psychologist dyadic interaction provides an opportunity to investigate interaction engagement under various settings introduced by differences in task conditions. In the study, we develop a computational system based solely on the observed vocal cues, specifically vocal prosody during the dyadic interaction. Furthermore, it is hypothesized that the engagement of the child can be predicted from the acoustic prosodic cues of both the child and the psychologist and data-driven methods can be designed to capture this relationship.

Our approach is inspired by several previous studies that link speech prosody to human behavior based constructs such as emotion (Austermann et al., 2005; Lee et al., 2011), approach-avoidance (Rozgic et al., 2011; Xiao et al., 2012), entrainment (Lee et al., 2014), blame/acceptance (Black et al., 2013), and empathy (Kempe, 2009; Aziz-Zadeh et al., 2010). These studies present techniques that computationally model prosodic patterns which are otherwise difficult to quantify perceptually. We build upon our previous work in Gupta et al. (2013, 2012) and present a data-driven approach to identify global prosodic patterns (task level statistics) as well as those that last over much shorter time spans (local

cues). The local level cues also provide a means for capturing the temporal relationship between the local prosodic patterns. We apply this model for engagement level prediction using prosody and our modeling technique can serve as a generic tool extendible to other modalities. We train separate models on the global and local cues and finally fuse their predictions. We observe that individual models using either the global or the local cues carry statistically significant predictive capability. We subsequently fuse the two components to utilize their complementarity. Our model assigns child engagement to one of the three instrument-defined categories, achieving an unweighted average recall of 55.8%. We also investigate the predictive capabilities of each individual cue used in classification.

Besides obtaining an objective decision, we also address the issues of variability introduced by interpersonal differences and dissimilar interaction settings. We show that our methodology captures prosodic patterns that are universally present across various interaction settings in the Rapid ABC protocol. Specifically, we normalize for individual speaker traits and train our model by combining data from all the tasks that comprise an entire dyadic interaction. We present the results categorized per task as well as for each interacting psychologist to estimate the generalizability of our model. We observe that our engagement model performs well over different parameters, but the performance does vary under different settings. This suggests that universal prosodic cues of engagement do exist, but they occur in combination with some setting specific patterns.

This paper is organized as follows: Section 2 describes the database. We explain the global and the local cue extraction scheme and the experimental setup in Section 3 and describe the classification approach and its results in Section 4. We present our analysis on the system in Section 5 and conclude in Section 6.

2. Database

We use the Rapid ABC database (Ousley et al., 2013; Rehg et al., 2013) collected at the Georgia Institute of Technology as part of an NSF Expeditions project. The Rapid ABC protocol is a 3-5 minutes long semi-structured interaction between a psychologist and a child during a predefined set of tasks. Concurrently, the psychologist assesses the child's social attention, non-verbal communication using gaze, vocalizations and facial expressions, and perceived engagement. The assessments by the psychologist are recorded on a paper form screener. Specifically, the recorded assessments in the screener are designed to identify behavioral markers of atypical social-emotional development, language and motor development. The primary purpose of this dataset is to aid experiments in designing technological solutions that would facilitate the integration of an autism screening tool into a medical office's work-flow. In the following sections, we describe the interaction settings for the Rapid ABC dataset, the screener form based evaluation followed by data statistics.

2.1. Interaction setup

The Rapid ABC interaction sessions involve semi-structured over-the-table interaction between a child and a psychologist. The session involves five tasks: (i) Smiling and saying hello, (ii) Ball play, (iii) Jointly looking at a book, (iv) Putting a book on the psychologist's head as if it was a hat, and (v) Smiling and tickling. These tasks are designed to capture various aspects related to cognitive, social, language and motor development.[1] The psychologist is provided a script for each task and concurrently records her assessments in the screener. The dataset is recorded using video, audio and physiology (wrist) sensors. In this work, we use the prosodic measures derived from the audio signal captured by a central farfield microphone on the table that records audio (at 16 khz sampling rate) from both the child and the psychologist.

2.2. Rapid ABC screener form

The Rapid ABC screener form was designed to concisely capture various observable/perceivable cues reflective of behavioral aspects like joint attention and social aptitude during each task in a session. The screener had to be manually completed by the psychologist while simultaneously interacting with the child. Part of the screener corresponding to the Ball play task is shown in Fig. 1. For each of the five tasks, the screener consists of two separate fields:

[1] A description of these tasks can be found at https://www.youtube.com/watch?v=89KnHRLz7EQ.

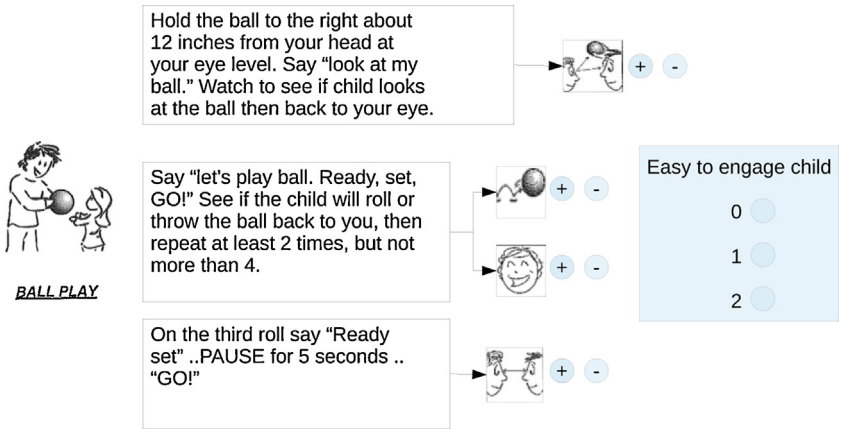

Fig. 1. Portion of the screener form for the ball play task.

Table 1
Instructions to the psychologist for annotating the engagement level.

Engagement level	Instruction to the psychologist
Level "0"(E_0)	Code a "0" if the interaction required no to little effort for you and if the child was ready and eager to engage. A "0" rating can be obtained even if some of the target behaviors were scored as not present, that is, as a 'minus'.
Level "1"(E_1)	Code a "1" if the interaction required some effort on the part of the examiner due to the childś shyness or distractability.
Level "2"(E_2)	Code a "2" if the interaction required extensive effort to engage the child. Also, if the child is highly fussy or refuses to interact, you would code a "2".

- Behavior annotation: The psychologist annotates a set of observed behaviors that are expected from the child during a task. These sets of behaviors of interest typically include actions reflecting joint attention or a social response. In Fig. 1, the instructions to annotate a set of behaviors for the ball task is provided. The psychologist scores a PLUS if behavior mentioned in the instructions is present and a MINUS otherwise.
- Ease of engagement: The psychologist annotates the child's engagement state as one of the 3 levels as per the instructions listed in Table 1. Whereas better metrics to quantify engagement may exist, the three levels used (classes) broadly stratify the engagement phenomenon making it easier for the psychologist to provide an objective judgment. A separate block to annotate the engagement levels (E) is provided in the screener form as shown in Fig. 1.

In our work, we aim to model the above perceived engagement level from observed cues, specifically, of speech prosody. Attributes related to behavioral annotation in the screener can be reliably annotated, but the engagement level annotation is prone to variabilities introduced by interpersonal differences amongst interacting psychologists; including in their perceptions, differences in settings under which the interaction happens, and the developmental stage of the child. We intend to aid the psychologist's judgment by minimizing the effect of the aforementioned factors.

2.3. Demographics and dataset statistics

We use 74 sessions containing speech recordings from 63 children in the Rapid ABC data. The Rapid-ABC protocol was repeated for 11 children as a follow up and hence we have two sessions for these children. The children are in the age range of 15–30 months and 39 of them are boys. These children interacted with one of the four psychologists trained for interaction in the Rapid ABC settings. Apart from the psychologist's assessments, these sessions contain manual lexical transcriptions with time alignments for child and psychologist vocalizations. The distribution of the engagement levels for each of the five tasks over the 74 sessions is listed in Table 2.

Table 2
Counts of the three levels of engagement (E_0, E_1, E_2) over the five tasks.

Task	Engagement level		
	E_0	E_1	E_2
Smiling and saying Hello (**Hello**)	55	13	6
Ball play (**Ball**)	62	7	5
Jointly looking at a book (**Book**)	47	18	9
Putting on a book on the head as if it is a hat (**Hat**)	71	1	2
Smiling and tickling (**Smiling**)	58	11	5
Total	293	50	27

From the table we observe that we have an uneven distribution of engagement levels. E_0 is the majority class in all the tasks. This suggests that the occurrence of other levels of engagement is rather atypical and possibly of greater interest with respect to the goals of the Rapid-ABC protocol in providing quick assessments over several aspects of child development. Also, the distribution of engagement levels varies depending on the task. For instance, the proportion of E_0 in the **Hat** task is significantly higher than any other task. Similarly, the proportion of E_1 in the **Book** task is significantly higher than **Ball, Hat** and **Smiling** tasks (We use a conservative difference in proportions test for p-value $<$ 5%. The number of samples for the significance testing is computed such that each class is considered to have the same number of samples as the least represented class.[2]). This indicates that the phenomenon of engagement is contingent on the task at hand and may vary under dissimilar settings.

3. Model development

We focus on using data-driven methods to model the psychologist's engagement level assignment using observed cues, specifically the vocal prosody of the participants. Even though the engagement level is conditioned upon the interaction settings, we hypothesize that there exist common vocal prosodic patterns reflective of engagement across these settings. Furthermore, we hypothesize that the child's engagement will be reflected not only in the child's prosodic cues, but also in those of the interacting psychologist. We aim to objectively capture these cues to infer the engagement levels. We construct multiple models performing the engagement evaluation and combine their decisions to obtain a final prediction. In the remainder of this section we describe our data preparation and prosodic cue extraction framework.

3.1. Dataset preparation

For each of the 74 sessions we have five tasks with engagement evaluation. We segment the audio files by task to allow for task-wise analysis. Hence we have 370 (74 sessions $\times$ 5 tasks) audio segments, each with corresponding manual diarization and an engagement score assigned by the interacting psychologist.

In order to train our models, we collectively use the 370 files from all the tasks and psychologists. Even though the data statistics suggest that engagement is contingent on the task, we combine the data primarily because of two reasons.

(i) First, we aim to capture prosodic patterns that are robust to the variability introduced by task-dependent contexts and different interaction partners.
(ii) Second, the small number of instances from E_1 and E_2 classes are not suitable for training specific models for each task and psychologist. As we rely on data-driven techniques, we need sufficient samples to reliably capture patterns.

[2] The rationale is to give equal importance to all the classes while performing the significance testing. This is particularly important as we use the unweighted average recall per class as our metric later. Since we reduce the number of samples, this test provides a more conservative significance level. However, we do avoid inflated significance which may arise due to different statistics on the majority class. For more details please refer to Bone et al. (2015).

After obtaining these separate audio chunks with an assigned engagement level $E \in \{E_0, E_1, E_2\}$, we extract the prosodic signals from the speech as discussed in the next section.

3.2. Prosodic signals

We extract a set of prosodic signals that characterize voice source activity: pitch, loudness, jitter and shimmer. All prosodic signals are computed using Praat (Boersma and Weenink, 2001) at a rate of 100 frames/s. In our experiments, we consider every 10 ms time interval as one analysis frame. Below, we detail prosodic signal computation including utilized signal denoising techniques.

- Speaker assignment: We use the manual segmentations to obtain a frame-wise speaker assignment vector $S = \{s_1, \ldots, s_n, \ldots, s_N\}$, where N represents the total number of analysis frames in the file (assignments are made every 10 ms). The element s_n assigns the nth frame of the audio file to either psychologist speech (Psy), child vocalization (Child), overlap (Ol) or silence (Sil).
- Pitch: We use an autocorrelation based method to perform pitch estimation (F_0, fundamental frequency) as described in Boersma (1993). We use an analysis window with a duration of 40 ms to estimate pitch at a time step of every 10 ms, which synchronizes with our speaker assignments. Since the extracted pitch may have errors due to audio quality, we smooth and cubically interpolate the pitch signal to reduce such errors. $P = \{p_1, \ldots, p_n, \ldots, p_N\}$ represents the vector of processed pitch values, where p_n is the pitch value for the nth frame.
- Intensity: We obtain the intensity estimates by squaring the audio magnitudes per frame. This is followed by convolution with a Gaussian window to reduce noise effects. The pitch-synchronous intensity ripple is also reduced by this operation to give a smoother intensity contour (Boersma and Weenink, 2001) (as convolution with Gaussian window is also a low-pass filtering operation). We represent the intensity vector as $I = \{i_1, \ldots, i_n, \ldots, i_N\}$.
- Jitter: Jitter serves as a measure of voice quality and is defined as the cycle-to-cycle variation of the fundamental frequency (F_0) (Farrús et al., 2007). We estimate the relative jitter using overlapping windows in the audio files. Relative jitter is computed by normalizing the absolute jitter (the average absolute difference between consecutive periods in speech signal) by the average period. Note that a jitter value for a specified window can only be calculated if it contains multiple F_0 value estimates. We chose a window length of one second shifted by 10ms. Smaller window lengths lead to several undefined jitter values and larger window lengths lead to imprecise estimates of local values of jitter as voicing from distant intervals is also incorporated.

As jitter measures cycle to cycle variation of periods in the speech signal, its estimation is sensitive to the accurate estimation of F_0. We can only estimate jitter when we have several pitch cycles in a window. Since we observed noisy jitter estimates in our data, we choose to smooth the jitter signal using a moving average filter. However, we do not interpolate the values as they are often missing over several windows, leading to poor interpolation. Over such windows with missing values, the jitter is listed to be undefined. We represent the set of jitter values as $J = \{j_1, \ldots, j_n, \ldots, j_N\}$, where j_n represents the jitter estimate for a window starting at the nth frame, extending for 1 s. Note that j_n can also be listed as undefined.

- Shimmer: Shimmer provides us with another measure of voice quality (Farrús et al., 2007). Shimmer measures cycle-to-cycle variation in intensity (jitter measured variation in periods). We estimate relative shimmer using the same window-wise approach as with jitter, smoothing but not interpolating the signal (again leading to undefined values). We represent the set of shimmer values as $H = \{h_1, \ldots, h_n, \ldots, h_N\}$, where h_n represents the shimmer estimate for a window starting at the nth frame.

3.3. Extracting cues from the prosodic signals

The characteristics of the prosodic signal based on engagement level can be captured using either (i) time series modeling tools or (ii) discriminative models on statistical measures computed over the prosodic signals. Time series modeling tools such as Gaussian mixture model-Universal background models (GMM-UBM) (Reynolds, 2002), i-vector systems (Dehak et al., 2009; Shum et al., 2010) train on frame-wise features and attempt to model the

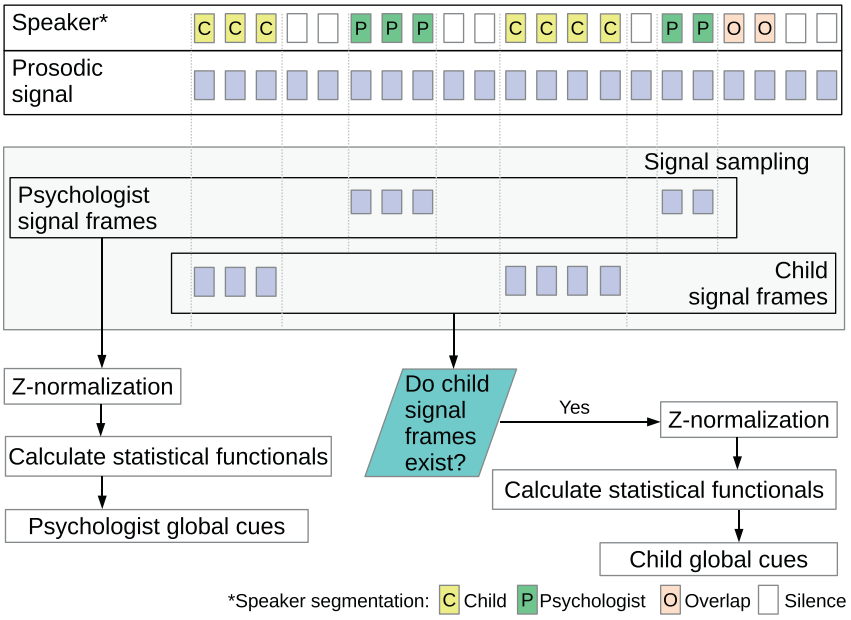

Fig. 2. Framework for global cues extraction.

probabilistic process governing the time series generation (conditioned on the class label). These methods assume that each sample in the time series is generated from a class dependent probability distribution. Given a set of time series from each class, these time series modeling tools estimate the parameters of the probability distribution for that class. On the other hand discriminative models on statistical estimates rely on capturing the differences between target classes using compact statistical representations. The latter is particularly useful in case of smaller datasets as modeling latent generative processes usually requires a large amount of data. Several other studies have also attempted to model similar time series data using compact statistical representations (Gupta et al., 2012; Bone et al., 2013; Hansen and Arslan, 1995). We use a similar discriminative scheme to capture the statistical properties of the prosodic time series at two levels of granularity: over the entire task duration (global cues) and at smaller time scales (local cues). The global cues help model the characteristics of the prosodic time series over the entire duration of a task, while the local cues quantify the local pattern in prosodic signals. We compare the outputs of the discriminative model based on local and global cues against a standard GMM-UBM model. We expect the discriminative model to perform better for two primary reasons, which are that the dataset is small and unbalanced and that the GMM-UBM model cannot account for temporal prosodic patterns. Over the next two section, we describe the local and global cues in detail.

3.3.1. Global cues

Global cues are statistical functional estimates calculated per-speaker over the entire interaction segment. Statistical functionals estimates computed over a sample set represent the characteristics of the underlying probability distribution from which the sample is drawn (Fernholz, 1983). These cues capture the overall characteristics but do not model the temporal evolution of the prosodic signals. Fig. 2 describes the extraction procedure and below we provide a step-wise description of our methodology involving speaker-specific signal sampling, normalization and statistical computation.

(i) Speaker-specific prosodic signal sampling: Given the speaker assignment S, we initially selectively sample the prosodic signals (pitch, intensity, jitter and shimmer) to contain frames only belonging to that speaker (overlaps excluded; this is shown in Fig. 2 in the block labeled "Signal sampling"). Note that in some of the tasks, we do not have child speech as the experimental design does not require a child to vocalize during interaction. For such tasks, we sample the prosodic signal corresponding to the psychologist speech only.

(ii) Speaker-wise signal normalization: Next, we perform speaker-wise normalization on the sampled signal values to minimize speaker specific traits. We chose z-normalization (similar to cepstral means–variance normalization

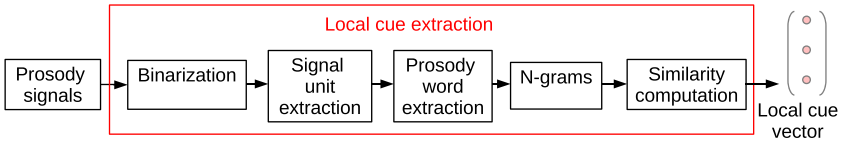

Fig. 3. Framework for local cue extraction.

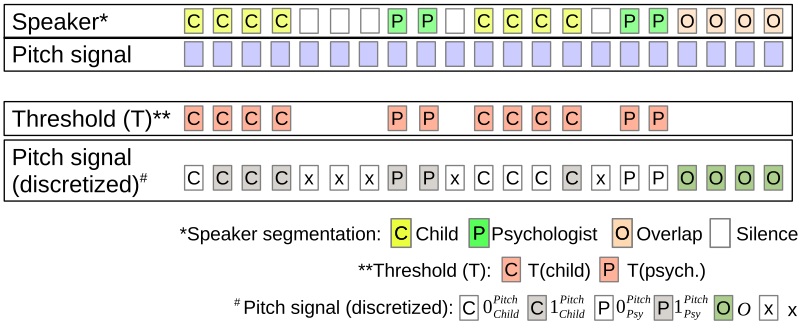

Fig. 4. Feature quantization representation. Please refer to Eq. (1) for the meaning of discretized pitch symbols.

(Molau et al., 2003)) with means and variances obtained from all available vocalizations for each speaker. This includes all the five tasks as well as audio recordings before and after the Rapid ABC sessions.

(iii) Global cue calculation: Finally, we calculate four statistical values over the z-normalized signals for the task at hand. These statistical functionals are mean, median, standard deviation and range. We call these our global cues.

Since the computation of global cues does not capture temporal dynamics associated with the interaction between the two participants, we propose a method to overcome these limitations in the following section, utilizing local prosodic cues.

3.3.2. *Local cues*

The global cues computation treats prosodic signals as time series of independently drawn samples and the cue value will be unaffected even if samples in the time series are interchanged. However, the temporal evolution of prosodic signals may impart further information regarding the engagement behavior of children. Therefore, we propose a "prosody word" based scheme to concisely capture the temporal patterns in prosody along with jointly modeling the prosody of the two speakers. The local cues quantify changes in prosodic signals (e.g. increase in intensity, decrease in pitch) which are later associated with the engagement levels. The local cues are inspired from feature quantization methods (Ahalt et al., 1990) and language modeling techniques (Katz, 1987) in automatic speech recognition (Levinson et al., 1983). Recently, feature quantization has been coupled with other modeling techniques to address problems such as topic modeling (Kim et al., 2012; Nakano et al., 2014) and speaker recognition (Shriberg et al., 2005). We evaluate the utility of these cues in engagement prediction over just using the global cues. Fig. 3 summarizes the local cue extraction framework, consisting of five steps which are described next in detail.

(i) Signal quantization: Given a prosodic signal and the speaker assignment, we quantize each frame of the signal based on a threshold (T). If the frame is assigned to a unique speaker (Child or Psy), we set T to be the median

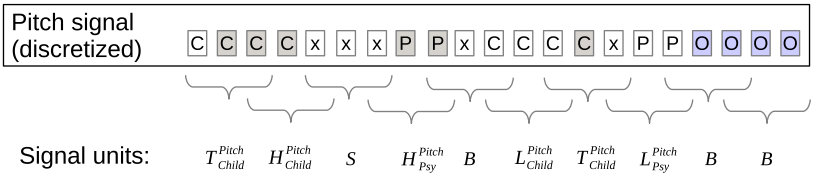

Fig. 5. A sample representation for signal unit extraction for the pitch signal.

Table 3
Feature unit assignment strategy for quantized signal values.

Signal Unit $U(D_n^{\text{Pitch}})$	Assignment strategy for the window D_n^{Pitch} ($D_n^{\text{Pitch}} = \{d(p_n), \ldots, d(p_{n+W})\}$)	
	Speaker present	Feature characteristics
$\text{H}_{\text{Psy}}^{\text{Pitch}}$	Only psyc. speech	discretized pitch value is **high** ($d(p_n) = 1_{\text{Psy}}^{\text{Pitch}}$) in the voiced frames
$\text{L}_{\text{Psy}}^{\text{Pitch}}$	”	discretized pitch value is **low** ($d(p_n) = 0_{\text{Psy}}^{\text{Pitch}}$) in the voiced frames
$\text{T}_{\text{Psy}}^{\text{Pitch}}$	”	There exists a **transition** from ($1_{\text{Psy}}^{\text{Pitch}}$) to ($0_{\text{Psy}}^{\text{Pitch}}$) or vice-versa
$\text{H}_{\text{Child}}^{\text{Pitch}}$	Only child speech	discretized pitch value is **high** ($d(p_n) = 1_{\text{Child}}^{\text{Pitch}}$) in the voiced frames
$\text{L}_{\text{Child}}^{\text{Pitch}}$	”	discretized pitch value is **low** ($d(p_n) = 0_{\text{Child}}^{\text{Pitch}}$) in the voiced frames
$\text{T}_{\text{Child}}^{\text{Pitch}}$	”	There exists a **transition** from ($1_{\text{Child}}^{\text{Pitch}}$) to ($0_{\text{Child}}^{\text{Pitch}}$) or vice-versa
B	Both speakers present	–
S	Window contains only silence	–

of the prosodic signal over the entire available vocal activity for the same speaker. As an example, a graphical illustration for pitch signal quantization is shown in Fig. 4. Eq. (1) lists the naming convention for quantized values in the provided example. $d(p_n)$ represents the discretized value for p_n (the pitch value at the nth frame). Frames assigned to overlap or silence are retained without further quantization.

$$d(p_n) = \begin{cases} 0_{\text{Psy}}^{\text{Pitch}} & \text{if } p_n < T(\text{Psy}); \text{if the frame belongs to the psychologistspeech} \\ 1_{\text{Psy}}^{\text{Pitch}} & \text{if } p_n \geq T(\text{Psy}); \text{if the frame belongs to the psychologist speech} \\ 0_{\text{Child}}^{\text{Pitch}} & \text{if } p_n < T(\text{Child}); \text{if the frame belongs to the child speech} \\ 1_{\text{Child}}^{\text{Pitch}} & \text{if } p_n \geq T(\text{Child}); \text{if the frame belongs to the child speech} \\ O & \text{if the frame contains an overlap} \\ \text{x} & \text{if the frame contains silence} \end{cases} \tag{1}$$

where T(Psy) is the median pitch value over the entire psychologist vocal activity and T(child) is the median pitch over the entire child vocal activity.

Although discretization reduces information, it allows for more complex modeling and learning with limited data. The feature median gives us a balanced distribution of the two binary categories and is not vulnerable to outliers. We do not perform a finer quantization of the prosodic signals as the number of prosody words (defined later) increases exponentially leading to sparsity issues.

(ii) Signal units: Next, we define a “signal unit” over a window consisting of multiple discretized signal frames, aiming to capture the signal dynamics over a shorter time span. We operate sequentially on the discretized prosodic signal given the window length W and window overlap length V. For a window starting at the nth frame, we define the signal unit based on the following W discretized signal values. A signal unit provides a compact statement about the prosody within a window, such as the window containing “high pitch” or “a transition from high to low intensity”. Using the same example of the pitch signal, we list the signal unit assignment strategy in Table 3. D_n^{Pitch} is a window starting at the nth frame containing the discretized values $\{d(p_n), \ldots, d(p_{n+W})\}$. $U(D_n^{\text{Pitch}})$ is the signal unit assigned to D_n^{Pitch}. An example pitch units assignment with $W = 3$ and $V = 1$ is shown in Fig. 5.

(iii) Prosodic words: Next, we concatenate the sequence of signal units from multiple signals to obtain the joint representation over various prosodic signals. In this work, we use the pitch and the intensity signals. Using more signals leads to an exponential increase in the number of prosodic words [3] and also jitter and shimmer are poorly

[3] Just adding one more signal increases the count of potential prosodic words from 6 to 29. This leads to a very large number of n-grams as computed next.

estimated [4]. Eq. (2) shows the vector obtained after concatenating signal unit for pitch $U(D_n^{Pitch})$ and intensity $U(D_n^{Int})$, defined as the prosodic word. The prosodic word provides a combined representation of the dynamics captured by the signal units

$$R_n = \begin{bmatrix} U(D_n^{\text{Pitch}}) \\ U(D_n^{\text{Int}}) \end{bmatrix} \tag{2}$$

Each of the prosodic words give us a crude estimation of the prosodic signal dynamics over a shorter time span.

(iv) N-grams of prosodic words: Given the sequence of prosodic words for the Rapid ABC tasks, we apply ideas similar to language modeling (Katz, 1987) in automatic speech recognition. We define n-grams on the set of prosodic words. For instance, the bigrams are defined by the pairs of consecutive prosodic words; R_n and $R_{(n+W-V)}$ (note that if a window starts at n, the next window will start at $n+W-V$ to achieve an overlap of V).

(v) Similarity computation: Next, we compute empirical occurrence probabilities of n-grams on: (a) a given test task and (b) all the tasks in the training set with a specified engagement level. Let n-gram$_l$ be one of L possible n-grams; then the empirical probability for n-gram$_l$ on the test task π_{test}(n-gram$_l$) is as shown in Eq. (3), and the empirical probability on training tasks with engagement level E, $\pi_{\text{train:}E}$(n-gram$_l$) are as shown in Eq. (4).

$$\pi_{\text{test}}(\text{n-gram}_l) = \frac{\text{Total count of } n-\text{gram}_l \text{ in the test task}}{\text{Total number of } n-\text{grams in the test task}} \tag{3}$$

$$\pi_{\text{train:}E}(\text{n-gram}_l) = \frac{\text{Total count of } n-\text{gram}_l \text{ in training partition tasks w/ engagement level } E}{\text{Total count of all } n-\text{grams in the training partition tasks w/ engagement level } E} \tag{4}$$

A final step in local cue computation consists of computing the cosine similarity C_E (Eq. (5)) between vectors of empirical probabilities π_{test}(n-gram$_l$) and $\pi_{\text{train:}E}$(n-gram$_l$). In Eq. (5), $\langle \pi_{\text{test}}(\text{n-gram}_1), \ldots, \pi_{\text{test}}(\text{n-gram}_L)\rangle$ represents the vector of π_{test}(n-gram$_l$) over all the n-grams computed on the test set. Similarly, $\langle \pi_{\text{train:E}}(\text{n-gram}_1)$, *ldots*, $\pi_{\text{train:E}}(\text{n-gram}_L)\rangle$ represents the vector of $\pi_{\text{train:}E}$(n-gram$_l$) computed on the train set. C_E is shown as the cosine distance between these two vectors. c_E(n-gram$_l$) is simply the product of π_{test}(n-gram$_l$) and $\pi_{\text{train:E}}$(n-gram$_l$). The C_E and c_E(n-gram$_l$) measures serve as our local cues and concisely capture the dynamics in prosodic signals. This last step in local cue computation is shown in Fig. 6.

$$C_E = \frac{\langle \pi_{\text{test}}(\text{n-gram}_1), \ldots, \pi_{\text{test}}(\text{n-gram}_L)\rangle^T}{|\langle \pi_{\text{test}}(\text{n-gram}_1), \ldots, \pi_{\text{test}}(\text{n-gram}_L)\rangle|_2} \quad \frac{\langle \pi_{\text{train:E}}(\text{n-gram}_1), \ldots, \pi_{\text{train:E}}(\text{n-gram}_L)\rangle}{|\langle \pi_{\text{train:E}}(\text{n-gram}_1), \ldots, \pi_{\text{train:E}}(\text{n-gram}_L)\rangle|_2} \tag{5}$$

$$c_E(\text{n-gram}_l) = \pi_{\text{test}}(\text{n-gram}_l) \cdot \pi_{\text{train:E}}(\text{n-gram}_l) \tag{6}$$

We train separate classifiers on the set of global and local cues as described in the next section. For the local cues, we tune the window parameters W and V using inner cross-validation on the training set. Given the small amount of data, only unigrams and bigrams are extracted for reliable estimation. As the number of n-grams is still large, we perform feature selection on c_E(n-gram$_l$) during classification via the correlation-based feature selection (CFS) filter algorithm proposed in Hall (1998). This feature selection scheme evaluates the worth of a subset of products c_E(n-gram$_l$) by considering the predictive ability of each along with the correlation amongst all of them. In the next section, we provide the results.

[4] The jitter and shimmer signals rely upon accurate estimation of F0. In our case, we find that jitter and shimmer estimates for psychologist and child speech are completely absent for about 33% and 50% of the 370 tasks, respectively, as F0 could not be continuously estimated by Praat over long periods of time in these sessions.

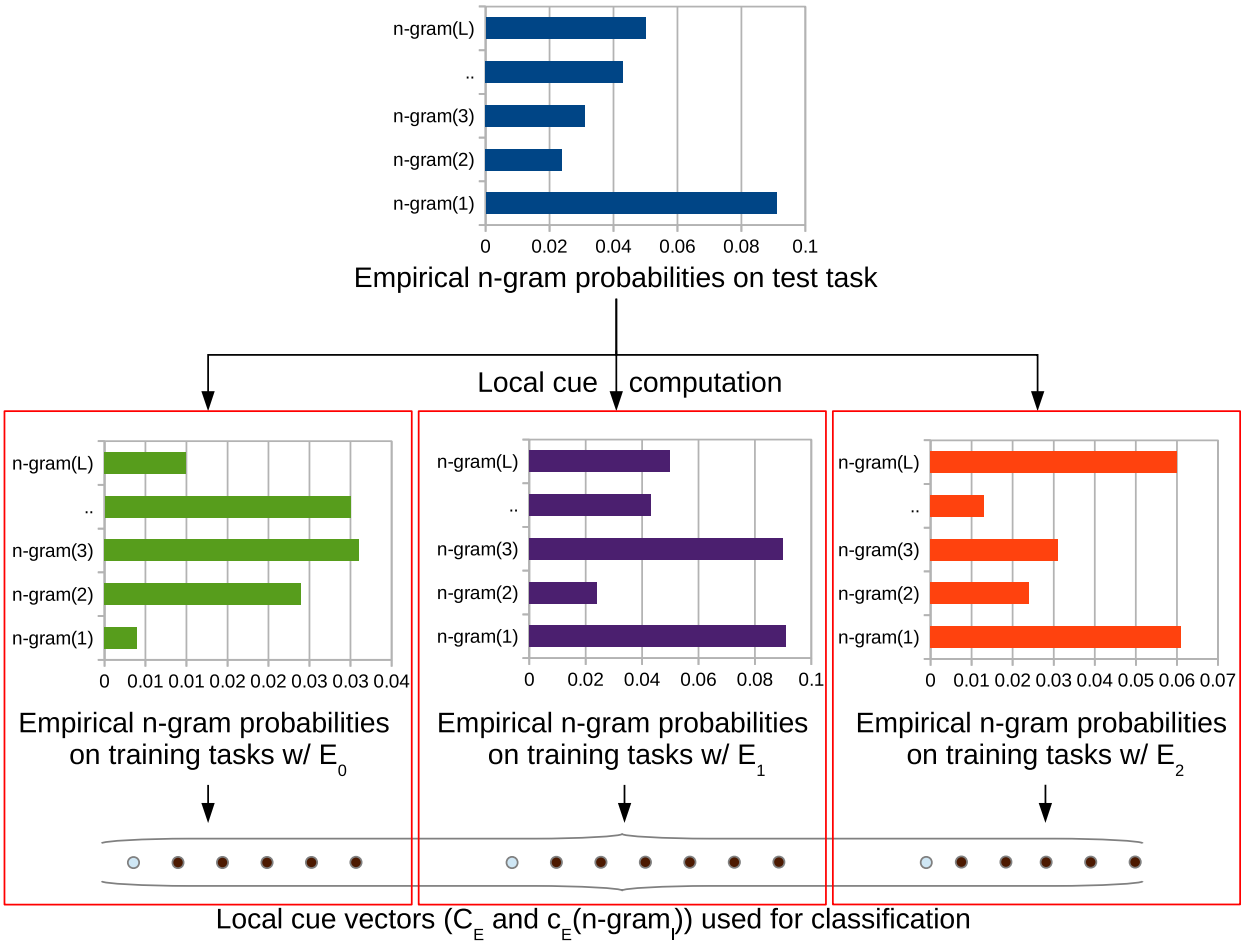

Fig. 6. Local cue computation between test task and training tasks. Blue dots represent cosine similarity (C_E, Eq. (5)) and the brown dots represent per-n-gram products (c_E(n-gram$_l$), Eq. (6)). C_E and c_E(n-gram$_l$) quantify the prosody signal dynamics and are used as features for engagement level classification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

4. Classification experiments

We initially train a GMM-UBM model on frame-wise features as a baseline. This is followed by the description of the discriminative model trained on the global and local cues. We first describe the GMM-UBM model followed by the discriminative model based on global and local cues.

GMM-UBM on frame-wise features: In this modeling scheme, we initially sample the prosodic signals belonging to the psychologist and the child speech frames (for speaker specific sampling please refer to Fig. 2). On the sampled prosodic signals, we train separate psychologist and child speech UBMs. These UBMs are then adapted using the data from tasks with specific engagement levels. For example, frames from psychologist speech belonging to all tasks with engagement level E_0 are used to adapt psychologist UBM to obtain a psychologist GMM-model for level E_0. Therefore we obtain three different GMM-models each for the two speakers, corresponding to the three engagement levels. Note that these models are trained on a 4-dimensional space defined by the frame-wise values of the four prosodic signals. We evaluate the GMM-UBM by using speaker-independent cross-validation, i.e., leaving sessions from one child for testing and training on the rest. On the test set, we evaluate the prosodic signals belonging to psychologist speech frames using psychologist GMMs (likewise for child frames). The final class likelihoods are obtained by summing up likelihoods from psychologist and child GMMs for that class. Given the unbalanced dataset, we use unweighted average recall (UAR) as our performance metric and the results using the GMM-UBMs are shown in Table 4.

Table 4
Classification results on GMM-UBM trained on the four frame-wise prosodic features.

UAR	Class recalls		
	E_0	E_1	E_2
38.0	48.5	36.0	29.6

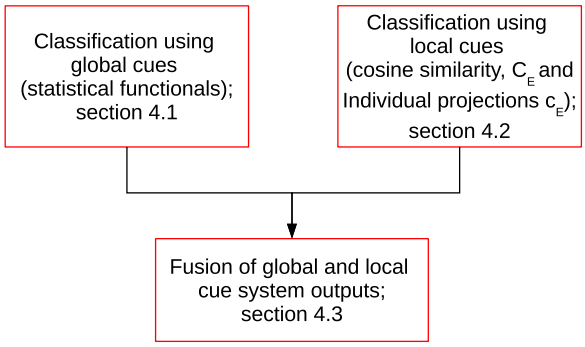

Fig. 7. Representation of the stacked generalization scheme used for classification. Separate classifiers are trained on global and local cues and finally outputs from these systems are fused.

From the results, we observe that though we beat the chance recall (33.33%) the value is relatively low. This is expected as we use the raw feature values without any regards to their characteristics over the entire task duration. Moreover in this simplistic model, the sequence of feature frames is not accounted for and no interaction between child and psychologist prosody dynamics is captured. We expect to represent the global characteristics and local temporal patterns using the global and local cues as presented in the next section.

Discriminative model based on global and local cues: We train individual classifiers on both global and local prosodic cues and then fuse the outputs in a stacked generalization framework (Wolpert, 1992). We chose this multi-layered classification approach for two main reasons. First, separate evaluations for global and local cues provide us with an independent measure of their discriminative power. Additionally, their joint performance helps us evaluate the degree of complementarity between the feature sets. Second, an independent training of global and local cue classifiers helps address data sparsity issues. A general schematic of classification experiments and their presentation in this section is given in Fig. 7.

Speaker-independent cross-validation is performed by holding out data segments for each child, leading to 63 splits. Given that the data suffers from class bias, we subsample data points from the majority class (E_0, E_1) so that each class has the same number of training instances as the least represented class (E_2). This effectively optimizes our performance metric unweighted average recall (UAR) as using the same number of samples per class in training assigns equal importance to individual class recalls.

In order to determine class boundaries, we train multiclass support vector machine (SVM) classifiers (Cortes and Vapnik, 1995) with pairwise boundaries. We obtain probabilistic decisions for each engagement level by fitting logistic models to data point distances from SVM hyperplanes. However, the logistic models may not yield class probabilities that sum to one. Thus these probabilities are scaled using the coupling method suggested by Hastie and Tibshirani (1998). The assigned class is the one with the highest probability. The parameters (i. e. kernel, boxconstraint) of the SVM classifier are tuned by internal cross-validation on the training set. We observe that all SVM classifiers perform best using a linear kernel as complex kernels may overfit a small dataset easily.

4.1. Classification using the global cues

We train two classifiers with global cues derived from the psychologist (SVM_{Psy}) and the child (SVM_{Child}). Whereas the psychologist speech is present in all the tasks, the child features are available only in 221 of 370 tasks (172 marked as E_0, 34 marked as E_1 and 15 marked as E_2). If the child speech is present, we fuse the outputs from SVM_{Psy} and

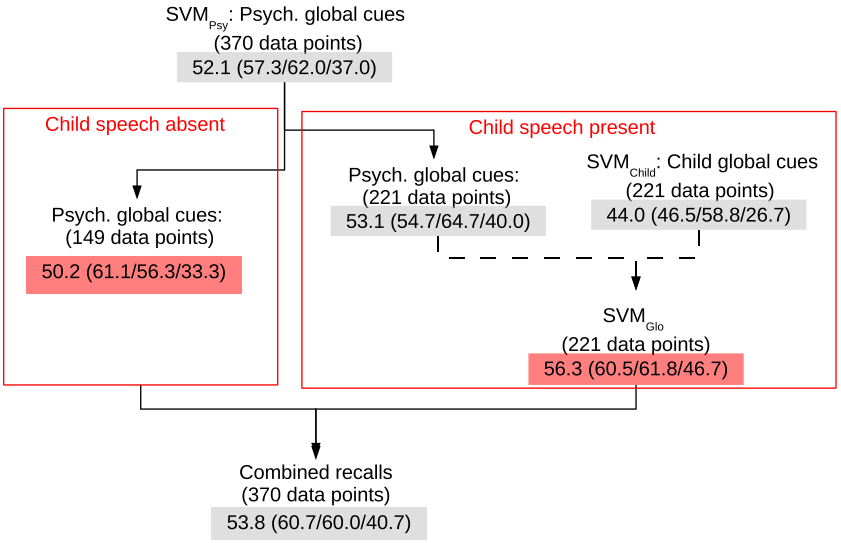

Fig. 8. Classification results using the global cues as features. Numbers represent the UAR across the three engagement levels with individual $E_0/E_1/E_2$ recalls in brackets.

SVM_{Child} using another classifier: SVM_{Glo}. We train SVM_{Glo} on the probabilities SVM_{Psy} and SVM_{Child} output on the training set itself. In the absence of child speech, we directly use the probabilities output from SVM_{Psy} to determine the overall UAR for the global system.

4.1.1. Results and discussion

A schematic of the global cue classification procedure and corresponding results (UAR and class recalls) are displayed in Fig. 8.

Global cues lead to statistically significant classification (53.8%, $p < 0.05$) compared to chance UAR (33.3%), supporting our hypothesis that the vocal prosody of speakers is related to the perceived engagement phenomenon. We observe that the features from psychologist speech are more predictive of the engagement levels than those from the child. This may be due to the fact that we train SVM_{Psy} over more training instances. Another contributing factor may be the fact that our database includes young children in early phases of language development. Hence child prosody may not hold as much information as the child may lack precise control and use of their voice source. However, we do observe some complementarity in predictive power from the child speech and the psychologist speech. Specifically, when child speech is present, fusion of SVM_{Psy} and SVM_{Child} improved UAR by 3.2% (absolute improvement) over SVM_{Psy} alone.

From the class-wise results, we observe that the classification recall for E_2 is the poorest. This may suggest that the global cues provide a better indication of the higher level of engagement and that information regarding the most disengaged level is diluted.

4.2. Classification using the local cues

The proposed local cues concisely capture the joint evolution of intensity and pitch from both the speakers. We expect the local cues to be complementary in information to the global cues as global cues do not model local events. We train an SVM classifier SVM_{Loc} on the local cues as features, i.e., the cosine similarities C_E and individual projections $c_E(\text{n-gram}_l)$ selected using the CFS filter algorithm.

4.2.1. Results and discussion

Classification results with local cues are shown in Fig. 9. After classification using SVM_{Loc}, UAR for instances with and without child speech are separately shown. This helps us interpret the gains made after fusing local and global cues in the absence/presence of child speech.

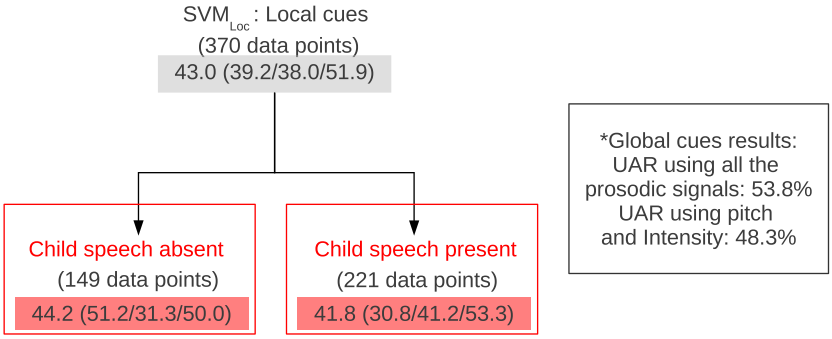

Fig. 9. Classification results using the local cues as features. Numbers represent the UAR across the three engagement levels with individual $E_0/E_1/E_2$ recalls in brackets. Results using global cues from all prosodic signals as well as pitch and intensity only are also shown for comparison.

We observe that characterization of prosodic signal dynamics performs marginally better (p-value $<.10$) than chance. The UAR is lower compared to the results using global cues, and the reasons may include: (i) use of only two prosodic signals against four used to extract global cues. A separate experiment on using global cues just from intensity and pitch give UAR of 48.3% against 53.8% obtained in Fig. 8. This is indicative of the loss in information after dropping jitter and shimmer. (ii) Extraction of local level cues involves discretization leading to loss of information. (iii) We do not have enough training samples from the minority classes to obtain a reliable empirical estimation of n-gram occurrence probabilities.

The global and local cues characterize prosodic signals at different temporal granularities. In the next section, we fuse the outputs from the two systems and investigate the complementarity between the local cues and global cues.

4.3. Fusion of global and local cues

We fuse the results from the sets of global and local cues using a final level of SVM classifiers. We train separate classifiers to fuse local and global system outputs as per the absence/presence of child speech. The $\text{SVM}_{\text{Fuse},C0}$ represents classifier trained on instances with no child speech and $\text{SVM}_{\text{Fuse},C1}$ in the presence of child speech. We cannot train a single fusion classifier as the global cue outputs are obtained from either SVM_{Psy} (child speech absent) or SVM_{Glo} (child speech present). Fig. 10 summarizes the results.

Our final model significantly beats the baseline GMM-UBM (using difference in proportions test for p-value $<5\%$. The number of samples are determined based on the conservative significance testing stated in Section 2.3) and the results are indicative of the degree of complementarity between local and global cues. An absolute gain of 2% is obtained over the model using global cues only. This indicates that the local cues, despite being weak individually, provide an extra source of information. A higher recall for E_1 and E_2 during the presence of child speech indicates that the cues on child speech favor classification toward lower engagement levels.

5. Global and local cue analysis

In this section, we further study the individual global and local cues through feature ranking and selection. We present our analysis separately for the two cases below.

5.1. Global cues

We use a one-R classifier (Holte, 1993) to rank the global cues obtained from each speaker over a dataset subsampled for balanced class distribution. A one-R classifier predicts the target label using only one feature at a time and this feature ranks the features based on their performance. We list the top five ranked features from the psychologist and the child in Table 5. We also plot the top two features in Fig. 11 to observe the inter-class patterns in a 2-dimensional space. Note that a separate classification experiment using the top few global features based on one-R ranking scheme did not improve the recall.

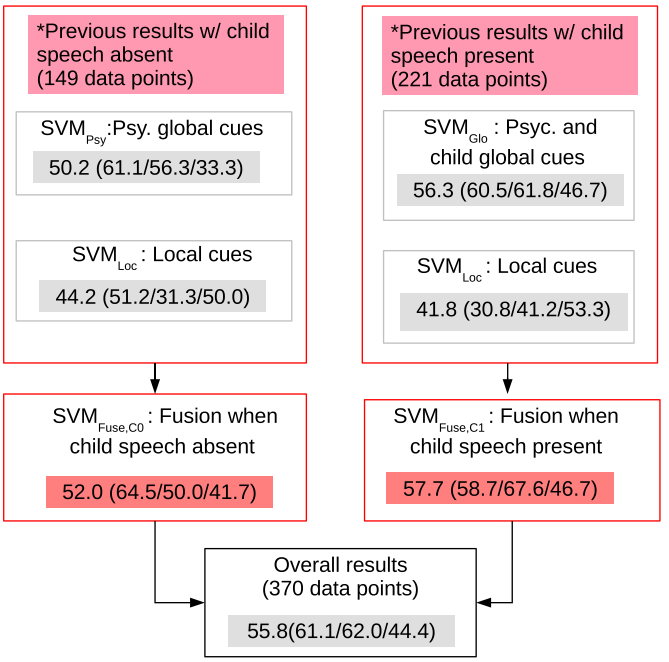

Fig. 10. Classification results after fusion. Numbers represent the UAR across the three engagement levels with individual $E_0/E_1/E_2$ recalls in brackets.

Table 5
List of top five global cues as ranked by the one-R classifier.

Psychologist	Child
Range (Pitch)	Mean (Jitter)
Range (Shimmer)	Mean (Pitch)
Median (Pitch)	Median (Jitter)
Median (Intensity)	Standard deviation (Intensity)
Standard deviation (Jitter)	Median (Shimmer)

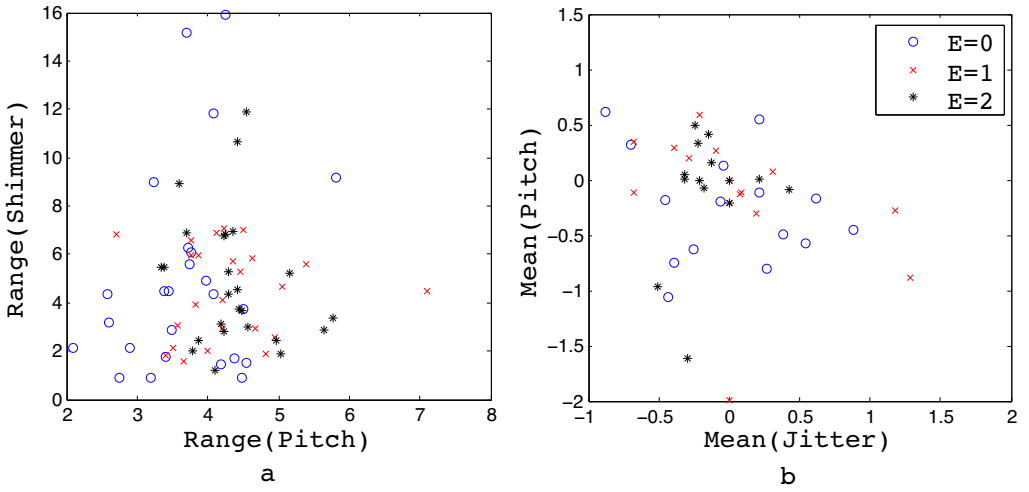

Fig. 11. Plot of the top two ranked global cues according to one-R classifier from (a) Psychologist (b) Child.

Table 6
Selection frequency for top 6 n-grams based on CFS filter algorithm. The representation of prosodic words is shown in Eq. (2). Uni-grams are represented as standalone prosodic words and bi-grams are shown as prosodic words in pairs.

N-gram and its selection frequency (out of 63 cross-validation folds)					
$\begin{pmatrix} H^{\text{Pitch}}_{\text{Child}} \\ H^{\text{Int}}_{\text{Child}} \end{pmatrix}$	$\begin{pmatrix} H^{\text{Pitch}}_{\text{Psy}} \\ T^{\text{Int}}_{\text{Psy}} \end{pmatrix}$ $\begin{pmatrix} H^{\text{Pitch}}_{\text{Psy}} \\ H^{\text{Int}}_{\text{Psy}} \end{pmatrix}$	$\begin{pmatrix} T^{\text{Pitch}}_{\text{Child}} \\ L^{\text{Int}}_{\text{Child}} \end{pmatrix}$ $\begin{pmatrix} L^{\text{Pitch}}_{\text{Child}} \\ T^{\text{Int}}_{\text{Child}} \end{pmatrix}$	$\begin{pmatrix} H^{\text{Pitch}}_{\text{Psy}} \\ T^{\text{Int}}_{\text{Psy}} \end{pmatrix}$ $\begin{pmatrix} L^{\text{Pitch}}_{\text{Child}} \\ T^{\text{Int}}_{\text{Child}} \end{pmatrix}$	$B\begin{pmatrix} L^{\text{Pitch}}_{\text{Child}} \\ T^{\text{Int}}_{\text{Child}} \end{pmatrix}$	$\begin{pmatrix} L^{\text{Pitch}}_{\text{Psy}} \\ H^{\text{Int}}_{\text{Psy}} \end{pmatrix}$
63	62	62	61	60	12

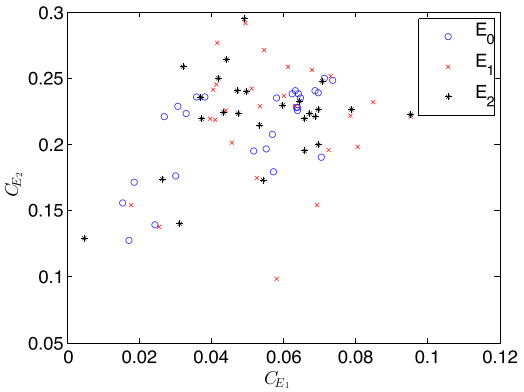

Fig. 12. Cosine similarities C_{E_1} and C_{E_2} for data subsampled to contain same number of instances per class.

A different set of features is selected from the psychologist and child global cues to be most predictive of engagement. A good mix of statistical functionals from different prosodic features suggests that the engagement phenomenon affects different aspects of vocal prosody. The plot of the top two features in Fig. 11 shows the class distribution between the two features. Between these features, there is no clear discrimination amongst classes. However we do observe a wider spread for instances from class E_0. Even though the class boundary characteristics can be different considering all the features together, a larger training can help better class boundary estimation and application of a more complex modeling technique.

5.2. Local cues

In this section, we analyze the discriminative power of the local cues, the cosine distance and the individual n-gram products. During each cross-validation fold, a subset of the n-gram products (c_E(n-gram$_l$)) is selected based on the CFS algorithm. Table 6 lists the selection frequency of the top seven n-grams for which a c_E(n-gram$_l$) is selected. We observe that certain n-grams get selected in almost all the cross-validation folds, suggesting some prosodic patterns are more informative than others. Also, as most of the selected n-grams are bi-grams, the count of a sequence of prosodic words is more important than stand alone uni-grams.

We also plot the cosine distances C_{E_1} and C_{E_2} over the dataset, sampled with an equal number of instances per class in Fig. 12. Similar to the global cues, in the selected two dimensional space, data points from the three classes overlap in the feature space. A higher value for cosine similarities C_{E_1} and C_{E_2} for test tasks assigned to E_1 and E_2 suggests that similar prosodic patterns exist amongst tasks with a specific engagement level. This encourages us to further investigate and improve our projection scheme and achieve a higher discrimination.

6. Effect of task type and interacting psychologist

We aim to derive robust universal measures reflective of engagement across variable interaction settings. We present an analysis of the performance of our system in relation to two sources of variability in our dataset: (i) the task at hand and (ii) the interacting psychologist.

Table 7
Task-wise performance of our classification scheme using prosodic features. Numbers in bracket show the true number of instances in each class.

Task	UAR	Class-wise Recall		
		E_0	E_1	E_2
Hello	54.9	54.5 (55)	76.9 (13)	33.3 (6)
Ball	66.9	54.8 (62)	85.6 (7)	60.0 (5)
Book	50.9	36.2 (47)	61.1 (18)	55.6 (9)
Hat	46.7	90.1 (71)	0.0 (1)	50.0 (2)
Smiling	38.3	58.6 (58)	36.4 (11)	20.0 (5)

Table 8
Results obtained per psychologist.

Psychologist #	No. of tasks annotated $E = (0 + 1 + 2)$	UAR	Class-wise Recall		
			E_0	E_1	E_2
1	200 (159 + 27 + 14)	59.6	58.5	70.3	50.0
2	130 (102 + 17 + 11)	48.4	61.8	47.1	36.4
3	25 (19 + 5 + 1)	44.6	73.7	60.0	0.0
4	15 (13 + 1 + 1)	89.7	69.2	100.0	100.0

6.1. Performance analysis per task

The observed distribution of engagement levels across subjects depends on the task, as previously shown in Table 2. We have thus far disregarded the task type, focusing on capturing more universal prosodic patterns. In this section, we investigate the task-wise performance of our approach. We split the results in Fig. 8 based on task type and list them in Table 7.

The recall of our system varies across each task. Performance is higher for **Hello**, **Ball**, and **Book** tasks compared to the lower performance in **Hat** and **Smiling** tasks (although all are above chance UAR, 33.3%). In particular, the E_1 recall for these three tasks is high, and these tasks account for 76% of the samples from E_1. We get a low UAR for the **Hat** task. However, given the high class imbalance, our systems performs well in correctly classifying 90.1% of the instances belonging to E_0. Our system performs the worst in the **Smiling** task. In this task, the psychologist approaches the child saying "I am gonna tickle you" three times and the child responds both visually and vocally, usually with a non-verbal vocalization like laughter or a grin. Hence the facial expression of the child may be a good indicator of his/her engagement in this case. Given the fact that we have a low UAR, we speculate that engagement may be better captured by a different modality than vocal prosody. Excluding the **Smiling** task, our prosody-based system achieves an UAR of 60.3% in the other four tasks, which is significantly higher than the performance of the **Smiling** task (using difference in proportions test for p-value $< 10\%$. The number of samples per class are determined based on the conservative significance testing stated in Section 2.3).

6.2. Performance analysis per psychologist

Each child is assessed by one of the four psychologists. Table 8 shows the performance of our system per psychologist. A higher UAR is obtained for psychologist #1 as compared to psychologist #2. This may suggest that our model has a slight bias towards the more represented psychologist. In particular, our model did not perform well for the minority classes for psychologist #2. It is hard to interpret the results for psychologist #3 and #4 given the small number of samples, but the model performs well for both these psychologists over the small representative set. Psychologist #3 has most of the samples assigned to E_0 and E_1 and the model achieves good class recalls for these classes. In the case of psychologist #4 we correctly classify 9 out of 13 instances from E_0 achieving good class recall. UAR is imbalanced for psychologist #3 and #4 as class recall for E_2 is computed on just one instance.

Overall, our model is able to capture the prosodic patterns that are universally utilized by the psychologists. As of now, the UARs are not significantly different (difference in proportions test) given this small set of samples per

psychologist. However, we intend to improve and analyze our framework in the future by including more samples, particularly from scarcely-represented psychologists.

7. Conclusion

Engagement behavior reflects a complex internal state signifying occupation with a person or in a task. Behavioral cues related to engagement may be conveyed visually, vocally or physiologically. In this work, we present a system to use the reflection of engagement in the audio modality, in particular in speech. We present our results over dyadic interaction between a child and a psychologist performing several interactive tasks together. We observe that the perceived engagement level of the child is not only reflected in the child's vocal prosody but also in the psychologist's prosody. This is particularly useful in the case when we predict the child's engagement in the absence of child speech. We develop a generic model to capture the patterns in a time series at two levels of temporal granularities and use it for finding patterns in prosody related to engagement behavior. The first component of our system uses global cues to capture the patterns over the entire duration of a task. In the second component, the local cues capture patterns which last over a shorter time span. The local cues are also capable of capturing the joint evolution of patterns across the two speakers, which is otherwise not accounted for by the global cues. We observe that the systems with global and local cues each provide discriminative power individually, but the best system is obtained after fusing their results. This suggests that the engagement phenomenon is related to the cross-sectional prosodic patterns projected over the entire interaction duration as well as the those over short durations.

We provide analysis on the derived patterns in the global and local cues which provide the discrimination on the engagement level. An analysis of performance per task suggests that our system works well in all tasks except the **smiling** task. Thus, our system generalizes fairly well under the variations introduced by a different kind of interaction setup. Similar observations were made across ratings from multiple psychologists, where our model was able to capture patterns used across different psychologists.

Our proposed classification schemes perform fairly well on the engagement prediction task from a single modality. We linked speech prosody to engagement phenomenon and the approach may be extended to other properties of speech. As we observe, our system performance may vary under a different interaction setting, other behavioral modalities beyond vocal prosody need to be included to model all observable patterns related to engagement, a goal for future enhanced system development. Our model, however, provides a general automatic behavior analysis tool using observed cues which when combined with the domain knowledge of the psychologist may provide an efficient child behavior assessment scheme.

Hence, our model provides a framework that may be applied to any such time series data but it can be further refined. Currently, the model is susceptible to sparsity problems and can benefit from smoothing techniques such as those common in language modeling (Chen and Goodman, 1999). Also we train a linear classifier to capture the patterns whereas the distribution of the top two features suggests a more clustered appearance. Hence a more efficient classification approach may be developed to take advantage of such a distribution. In addition, since our model captures the common patterns without regards to the factors of variability, an optimal combination with models specific to each setting may further improve the results. Our classifier combines the results from the two components and further investigation needs to be done to understand how the two cues complement each other.

Acknowledgment

This work is supported by National Science Foundation (grant number: NSF IIS-1029373).

References

Ahalt, S.C., Krishnamurthy, A.K., Chen, P., Melton, D.E., 1990. Competitive learning algorithms for vector quantization. Neural Netw. 3, 277–290.

Akhtar, N., Dunham, F., Dunham, P.J., 1991. Directive interactions and early vocabulary development: the role of joint attentional focus. J. Child Lang. 18, 41–49.

Austermann, A., Esau, N., Kleinjohann, L., Kleinjohann, B.,2005. Prosody based emotion recognition for mexi. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS 2005). IEEE, pp. 1138–1144.

Aziz-Zadeh, L., Sheng, T., Gheytanchi, A., 2010. Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. PLOS ONE 5, e8759.

Biller, H.B., 1993. Fathers and Families: Paternal Factors in Child Development. ABC-CLIO.
Black, M.P., Katsamanis, A., Baucom, B., Lee, C.-C., Lammert, A., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2013. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. Speech Commun. 55, 1–21.
Boersma, P., Weenink, D., 2001. Praat Tool. Institute of Phonetics Sciences of the University of Amsterdam.
Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the institute of phonetic sciences, vol. 17, Amsterdam, pp. 97–110.
Bone, D., Lee, C.-C., Chaspari, T., Black, M.P., Williams, M.E., Lee, S., Levitt, P., Narayanan, S., 2013. Acoustic-prosodic, turn-taking, and language cues in child–psychologist interactions for varying social demand.
Bone, D., Goodwin, M.S., Black, M.P., Lee, C.-C., Audhkhasi, K., Narayanan, S., 2015. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J. Autism Dev. Disord. 45, 1121–1136.
Casey, A.M., McWilliam, R., 2005. Where is everybody? organizing adults to promote child engagement. Young Except. Child. 8, 2–10.
Chen, S.F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. 13, 359–393.
Cielinski, K.L., Vaughn, B.E., Seifer, R., Contreras, J., 1995. Relations among sustained engagement during play, quality of play, and mother–child interaction in samples of children with down syndrome and normally developing toddlers. Infant Behav. Dev. 18, 163–176.
Cloward, R.A., Ohlin, L.E., Cloward, R.A., 1960. Delinquency and Opportunity: A Theory of Delinquent Gangs, 90559. Free Press, New York.
Coplan, J., Gleason, J.R., 1990. Quantifying language development from birth to 3 years using the early language milestone scale. Pediatrics 86, 963–971.
Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20, 273–297.
Davie, R., Butler, N., Goldstein, H., 1972. From Birth to Seven: The Second Report of the National Child Development Study.(1958 cohort)). Longmans, London, pp. 198, p. 1.
de Kruif, R.E., McWilliam, R., 1999. Multivariate relationships among developmental age, global engagement, and observed child engagement. Early Child. Res. Q. 14, 515–536.
Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Interspeech, volume 9, pp. 1559–1562.
Delano, M., Snell, M.E., 2006. The effects of social stories on the social engagement of children with autism. J. Posit. Behav. Interv. 8, 29–42.
Dodge, K.A., Lansford, J.E., Burks, V.S., Bates, J.E., Pettit, G.S., Fontaine, R., Price, J.M., 2003. Peer rejection and social information-processing factors in the development of aggressive behavior problems in children. Child Dev. 74, 374–393.
Egeland, B., Farber, E.A., 1984. Infant–mother attachment: factors related to its development and changes over time. Child Dev.
Farrús, M., Hernando, J., Ejarque, P., 2007. Jitter and shimmer measurements for speaker recognition. In: INTERSPEECH, pp. 778–781.
Fernholz, L.T., 1983. Von Mises Calculus for Statistical Functionals. Springer.
Göncü, A., 1999. Children's Engagement in the World: Sociocultural Perspectives. Cambridge University Press.
Gupta, R., Lee, C.-C., Bone, D., Rozga, A., Sungbok, L., Narayanan, S., 2012. Acoustical analysis of engagement behavior in children. In: Workshop on Child, Computer and Interaction, Portland.
Gupta, R., Lee, C.-C., Narayanan, S.,2012. Classification of emotional content of sighs in dyadic human interactions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2265–2268.
Gupta, R., Lee, C.-C., Sungbok, L., Narayanan, S., 2013. Assessment of a child's engagement using sequence model based features. In: Workshop on Affective Social Speech Signals, Grenoble.
Hall, M.A., 1998. Correlation-Based Feature Subset Selection for Machine Learning. University of Waikato, Hamilton, New Zealand, Ph.D. thesis.
Hansen, J.H., Arslan, L.M.,1995. Foreign accent classification using source generator based prosodic features. In: International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95, vol. 1. IEEE, pp. 836–839.
Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (Eds.), In: Advances in Neural Information Processing Systems, vol. 10. MIT Press.
Holte, R., 1993. Very simple classification rules perform well on most commonly used datasets. Mach. Learn. 11, 63–91.
Kasari, C., Gulsrud, A.C., Wong, C., Kwon, S., Locke, J., 2010. Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. J. Autism Dev. Disord. 40, 1045–1056.
Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing 35, 400–401.
Kempe, V., 2009. Child-directed speech prosody in adolescents: relationship to 2d: 4d, empathy, and attitudes towards children. Personal. Individ. Differ. 47, 610–615.
Kim, S., Georgiou, P., Narayanan, S.S., 2012. Latent acoustic topic models for unstructured audio classification. APSIPA Trans. Signal Inf. Process. 1.
Kishida, Y., Kemp, C., 2006. Measuring child engagement in inclusive early childhood settings: implications for practice. Aust. J. Early Child. 31, 14.
Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. 53, 1162–1171.
Lee, C.-C., Katsamanis, A., Black, M.P., Baucom, B., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2014. Computing vocal entrainment: a signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. Comput. Speech Lang. 28, 518–539.
Levinson, S.E., Rabiner, L.R., Sondhi, M.M., 1983. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst. Tech. J. 62, 1035–1074.
Libbey, H.P., 2004. Measuring student relationships to school: attachment, bonding, connectedness, and engagement. J. Sch. Health 74, 274–283.

Maher Ridley, S., McWilliam, R., Oates, C.S., 2000. Observed engagement as an indicator of child care program quality. Early Educ. Dev. 11, 133–146.

Manning, B.H., White, C.S., Daugherty, M., 1994. Young children's private speech as a precursor to metacognitive strategy use during task engagement. Discourse Process. 17, 191–211.

McWilliam, R., Casey, A.M., 2008. Engagement of Every Child in the Preschool Classroom. Paul H. Brookes Publishing Company.

McWilliam, R., Scarborough, A.A., Kim, H., 2003. Adult interactions and child engagement. Early Educ. Dev. 14, 7–28.

Molau, S., Hilger, F., Ney, H.,2003. Feature space normalization in adverse acoustic conditions. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03), vol. 1. IEEE, pp. I–656.

Morrissey-Kane, E., Prinz, R.J., 1999. Engagement in child and adolescent treatment: the role of parental cognitions and attributions. Clin. Child Fam. Psychol. Rev. 2, 183–198.

Nakano, T., Yoshii, K., Goto, M.,2014. Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5202–5206.

Nes, L.S., Segerstrom, S.C., Sephton, S.E., 2005. Engagement and arousal: optimism's effects during a brief stressor. Personal. Soc. Psychol. Bull. 31, 111–120.

Newborg, J., Stock, J.R., Wnek, L., Guidubaldi, J., Svinicki, J., et al., 1984. Battelle Developmental Inventory. DLM Teaching Resources, Allen, TX.

Ousley, O.Y., Arriaga, R.I., Morrier, M.J., Mathys, J.B., Allen, M.D., Abowd, G.D., 2013. Beyond parental report: findings from the rapid-abc, a new 4-minute interactive autism, Technical report series: report number 100. Center for Behavior Imaging, Georgia Institute of Technology (http://www.cbi.gatech.edu/techreports).

Poulsen, A.A., Ziviani, J.M., 2004. Can i play too? physical activity engagement of children with developmental coordination disorders. Can. J. Occup. Therapy 71, 100–107.

Read, J., MacFarlane, S., Casey, C., 2002. Endurability, engagement and expectations: measuring children's fun. Interaction Design and Children, vol. 2. Shaker Publishing Eindhoven, pp. 1–23.

Rehg, J.M., Abowd, G.D., Rozga, A., Romero, M., Clements, M.A., Sclaroff, S., Essa, I., Ousley, O.Y., Li, Y., Kim, C., et al.,2013. Decoding children's social behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3414–3421.

Reynolds, D.A., 2002. An overview of automatic speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075).

Rogers, S.J., 2000. Interventions that facilitate socialization in children with autism. J. Autism Dev. Disord. 30, 399–409.

Rozgic, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P.G., Narayanan, S.S., 2011. Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach. In: Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP).

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W., Paiva, A.,2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: 6th ACM/IEEE International Conference on Human–Robot Interaction (HRI). IEEE, pp. 305–311.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., 2005. Modeling prosodic feature sequences for speaker recognition. Speech Commun. 46, 455–472.

Shum, S., Dehak, N., Dehak, R., Glass, J.R., 2010. Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In: Odyssey, p. p. 16.

Skinner, E.A., Belmont, M.J., 1993. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. J. Educ. Psychol. 85, 571.

Skinner, E.A., Wellborn, J.G., Connell, J.P., 1990. What it takes to do well in school and whether i've got it: a process model of perceived control and children's engagement and achievement in school. J. Educ. Psychol. 82, 22.

Taylor, B.M., Pearson, P.D., Peterson, D.S., Rodriguez, M.C., 2003. Reading growth in high-poverty classrooms: the influence of teacher practices that encourage cognitive engagement in literacy learning. Element. Sch. J., 3–28.

Tomasello, M., Farrar, M.J., 1986. Joint attention and early language. Child Dev., 1454–1463.

Volkmar, F.R., Carter, A., Sparrow, S.S., Cicchetti, D.V., 1993. Quantifying social development in autism. J. Am. Acad. Child Adolesc. Psychiatry 32, 627–632.

Walker, S.P., Wachs, T.D., Meeks Gardner, J., Lozoff, B., Wasserman, G.A., Pollitt, E., Carter, J.A., 2007. Child development: risk factors for adverse outcomes in developing countries. Lancet 369, 145–157.

Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.

Xiao, B., Georgiou, P.G., Narayanan, S.S., 2012. Multimodal detection of salient behaviors of approach-avoidance in dyadic interactions. In: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2012).

Xu, X., Mellor, D., Kiehne, M., Ricciardelli, L.A., McCabe, M.P., Xu, Y., 2010. Body dissatisfaction, engagement in body change behaviors and sociocultural influences on body image among Chinese adolescents. Body Image 7, 156–164.

Yatchmenoff, D.K., 2005. Measuring client engagement from the client's perspective in nonvoluntary child protective services. Res. Soc. Work Pract. 15, 84–96.

Yu, C., Aoki, P.M., Woodruff, A., 2004. Detecting user engagement in everyday conversations, arXiv preprint cs/0410027.