Paralinguistic Analysis of Children's Speech in Natural Environments.

Hrishikesh Rao, Mark A. Clements, Yin Li, Meghan R. Swanson, and Daniel S. Messinger

Abstract Paralinguistic cues are the non-phonemic aspects of human speech that convey information about the affective state of the speaker. In children's speech, these events are also important markers for the detection of early developmental disorders. Detecting these events in hours of audio data would be beneficial for clinicians to analyze the social behaviors of children. The chapter focuses on the use of spectral and prosodic baseline acoustic features to classify instances of children's laughter and fussing/crying while interacting with their caregivers in naturalistic settings. In conjunction with baseline features, long-term intensity-based features, that capture the periodic structure of laughter, enable in detecting instances of laughter to a reasonably high degree of accuracy in a variety of classification tasks.

1 Paralinguistic Event Detection in Toddlers' Interactions with Caregivers

Paralinguistic cues are non-phonemic aspects of human speech that are characterized by modulation of pitch, amplitude, and articulation rate [**?**]. These cues convey

Hrishikesh Rao
Georgia Institute of Technology, Atlanta, GA, USA e-mail: hrishikesh@gatech.edu

Mark A. Clements
Georgia Institute of Technology, Atlanta, GA, USA e-mail: clements@gatech.edu

Yin Li
Georgia Institute of Technology, Atlanta, GA, USA e-mail: yli440@gatech.edu

Meghan R. Swanson
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA e-mail: meghan.swanson@cidd.unc.edu

Daniel S. Messinger
University of Miami, Coral Gables, FL, USA e-mail: dmessinger@miami.edu

information about the affective state of the speaker and can be used to change the semantic content of a phrase being uttered. For example, the phrase, “Yeah right”, when modulated with laughter indicates sarcasm [?]. Paralinguistic cues encompass the commonly produced ones such as crying and coughing to those that are widely considered to be social taboos such as belching and spitting [?].

Charles Darwin, in his seminal work on emotions in animals, described laughter as a paralinguistic cue used primarily to convey joy or happiness [?]. Laughter is a signal which consists of vowel-like bursts that has been found to be a highly variable signal. Adults produce laugh-like syllables, which are repetitive in nature and the production rates in laughter are higher than those of speech-like sounds [?]. Laughter also tends to have a higher pitch and variability compared to speech. Laughter is a socially rich signal that manifests itself in different forms. Laughter bouts have been classified as being “song-like” which consists of modulation of pitch, “snort-like” with unvoiced portions, and “unvoiced grunt-like” [?]. Furthermore, research has used laughter labels based on the type of stimulus used to produce it [?]. This includes joy, taunting, schadenfreude, and tickling. Although, laughter is considered to be a signal for indicating positive affect, the perception of laughter can change based on the context in which it is used. In speed dating situations, women were rated to be flirting if they laughed while interacting with men [?].

Paralinguistic cues, such as laughter and crying, play an important role in children’s early communication, and these cues are useful in conveying the affective state of the speaker. The cues have also been found to differ when infants and children with autism spectrum disorder (ASD) are compared to controls [?, ?]. The diarization of such events in extended recordings has shown preliminary evidence as a utility in the diagnosis detecting pathologies [?]. These events can also be used to analyze children’s communicative behaviors in social interactions with their caregivers. Laughter is primarily used to express positive affect and has been found to usually follow a state of anticipatory arousal, especially tickling [?]. Fussing/Crying could indicate that the child is upset or disinterested in the task being initiated by the caregiver in a dyadic setting.

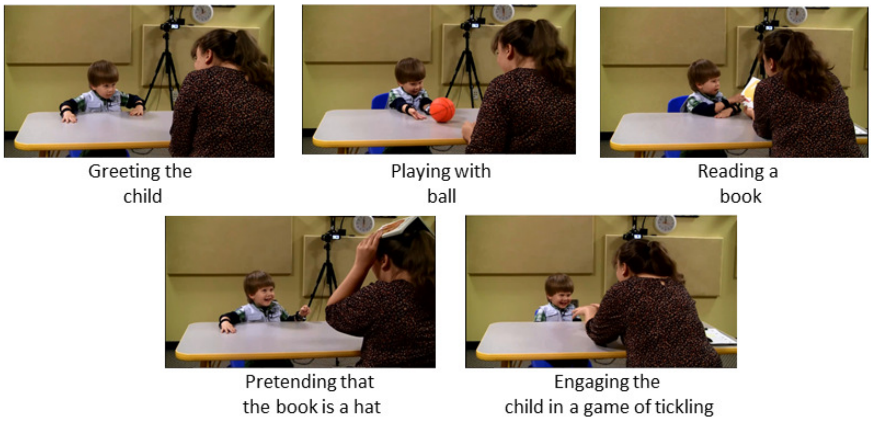
1.1 Databases

The research will focus on using long-term syllable-level features to detect laughter in children’s speech. For this purpose, three datasets will be used. For detecting laughter in children’s speech, we have used the MMDB, Strange Situation, and the IBIS dataset datasets.

1.1.1 Multi-modal Dyadic Behavior Dataset

The Multi-modal Dyadic Behavior (MMDB) dataset [?] consists of recordings of semi-structured interactions between a child and an adult examiner. The recordings are of multi-modal in nature and consists of video, audio, and physiological data. The sessions of the MMDB were recorded in the Child Study Lab (CSL) at the Georgia Institute of Technology, Atlanta, USA.

The protocol in this study is the Rapid ABC play protocol which is a short (3-5 minute) interaction between a trained examiner and a child whose interaction skills are assessed based on social attention, back-and-forth interactions, and nonverbal communication which have been indicative of socio-communicative milestones. The Rapid-ABC consists of five stages, which is illustrated in Figure **??**, and these consist of greeting the child by calling his or her name, rolling a ball back-and-forth with the child, reading a book and eliciting responses from the child, placing the book on the head and pretending it to be a hat, and engaging the child in a game of tickling.



Greeting the child

Playing with ball

Reading a book

Pretending that the book is a hat

Engaging the child in a game of tickling

Fig. 1 Stages of the dyadic interaction between child and examiner in the MMDB.

The annotations of the MMDB dataset were performed by research assistants in the CSL and were coded for the different stages of the Rapid-ABC protocol. For the speech modality, the child's vocalization events such as speech, laughter, and fussing/crying along with the examiner's transcribed speech events were annotated.

The database currently has recordings from 182 subjects with 99 males and 83 females (aged 15-29 months) and there were 54 follow up visits. The annotations of the social behaviors were performed using the open-source annotation tool ELAN and the screenshot of the ELAN software with the annotations for one of the MMDB sessions is shown in Figure **??**.

The dataset is significant in a multitude of ways, mainly from the fact that this represents one of the very few datasets available to the scientific community which has a rich variation in the number of subjects and the range of ages. From the speech perspective, there are vocalizations involving laughter and fussing/crying and are present in a significant number with most of the laughter samples emanating during the tickling stage of the Rapid-ABC. The child's vocalizations are recorded using

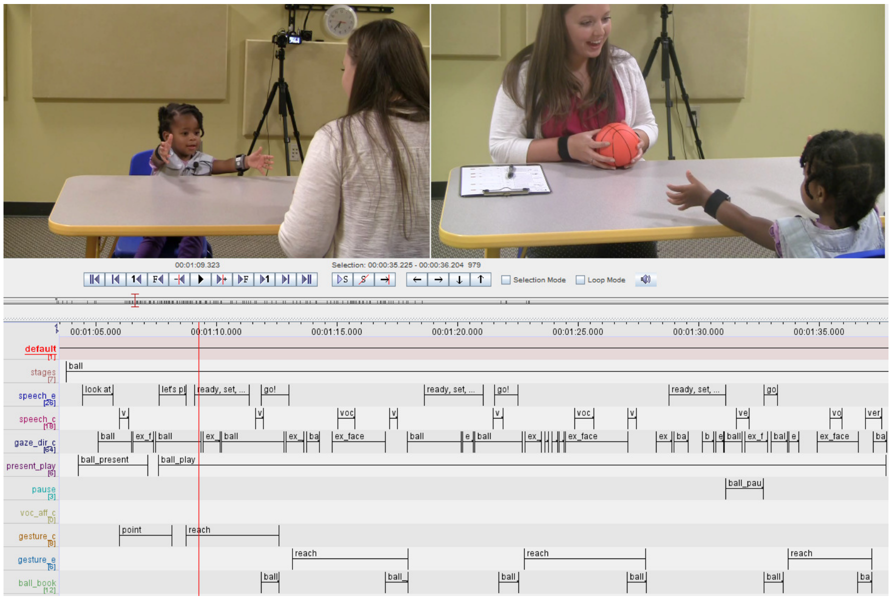


Fig. 2 MMDB session annotations in ELAN.

lavalier microphones which are in close proximity to the child and are generally free from any type of noise. From the multi-modal perspective, this dataset represents a challenging prospect to analyze the interaction of laughter and smiling in children and fuse information from audio and video sources to detect instances of laughter.

1.1.2 Strange Situation

The Strange Situation Procedure [**?**] is used for analyzing attachment behaviors of children with their caregivers. The strange situation protocol consists of eight episodes, each of which is three minutes in duration. In episodes 1–3, the child (in the company of the caregiver) is first confronted with a strange environment (a play room) and then with a stranger (an unknown research assistant). During the fourth episode, the caregiver leaves the room and the infant is left with the stranger. The caregiver returns during the fifth episode and the stranger leaves. The caregiver then leaves again (episode 6), which means the infant is alone in the room. The stranger returns (episode 7), and eventually the caregiver also returns(episode 8).

The stressful situations which elicit attachment behaviors in children include the environment in which the child is in, the stranger with whom the child is with, and the separations from the caregiver. The goal is to evaluate how the child reacts to being reunited with the mother, specifically, whether he/she approaches her, is soothed by the contact, and returns to play. Attachment behaviors with the caregiver on reunion lead to classifcation into one of three categories: secure, insecure avoidant, or insecure resistant. These attachment styles along with their prototypical crying patterns during reunion episodes are shown in Table **??** [**?**]. Crying is an important behavior in attachment classification from the Strange Situation Procedure.

Table 1 Classification criteria using crying in the Strange Situation Procedure for the three different attachment categories as described by Waters, 1978

Attachment behavior	Crying
Avoidant	Low (preseparation), high or low (separation), low (reunion)
Secure	Low (preseparation), high or low (separation), low (reunion)
Ambivalent	Occasionally (preseparation) , high (separation), separation) moderate to high (reunion)

The Strange Situation dataset analyzed in was provided by Daniel Messinger from research conducted at the University of Miami, Coral Gables, FL, USA. This dataset consists of strange situation recordings from 34 infants of 12 months of age and were recorded using the LENA device [**?**]. The annotations provided by the collaborators consists of child's speech, crying, and laughter. The dataset is beneficial from the point of view of testing models trained on the MMDB and testing it on the Strange Situation corpus. The importance of the dataset emanates from the fact that the recordings come from noisy conditions and the type of crying produced in the Strange Situation consists of wailing while that of the MMDB is more of whimpering in nature.

1.1.3 Infant Brain Imaging Study

The Infant Brain Imaging (IBIS) study is an ongoing longitudinal study of infants at high and low familial risk for ASD. The study includes [**?**, **?**] a dataset of recordings consisting of infants' speech which has been recorded in the homes of their caregivers and external environments such as grocery stores, playschools, and shopping malls. The IBIS study includes four clinical sites: University of North Carolina, Chapel Hill; University of Washington, Seattle; The Childrens Hospital of Philadelphia; and Washington University, St. Louis, and data coordination at Montreal Neurological Institute, McGill University. The current dataset includes a subsample of IBIS participants from the University of North Carolina and The Childrens Hospital of Philadelphia. Data was recorded at 9 and 15 months of age generating a total of 85 recordings. The distribution of the subjects based on their risk factors is shown in Table **??**.

The recordings of the infants's interactions with their caregivers are 16 hours in length and were recorded using the Language Environment Analysis (LENA) device

Table 2 Risk factor of ASD for the subjects in the IBIS study at 9 and 15 months of age.

	Low Risk	High Risk
9 months of age	16	37
15 months of age	7	25

which is a portable digital language processor. The LENA device is a light-weight audio recorder which can easily fit inside the vest worn by an infant. The recorder, shown in Figure **??**, has the ability to record single channel audio data at a sampling rate of 16 kHz.

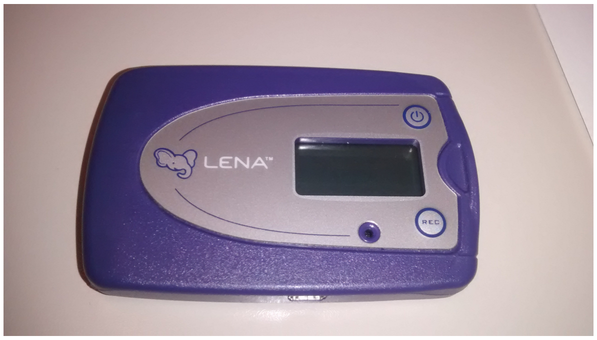


Fig. 3 LENA audio recording device used for infant vocal development analysis.

The software provided along with the recorder is a data mining tool, LENA Advanced Data Extractor (ADEX), which can potentially be useful for analyzing the various segments in day-long recordings. The tool has the capability of segmenting and parsing various information about the audio events of interest. These include the infant's and adult's vocalizations, cross-talk, background noise, electronic noise, and turn-taking events [**?**].

The LENA software does not provide a fine-grained analysis of the infant's nonverbal vocalizations and does not provide timestamps of when the infant laughed, cried, or produced other paralinguistic vocalizations. These important measures are potentially valuable in understanding the social behaviors of infants when they interact with their caregivers.In the context of infants at high-risk for ASD, the atypical characteristics of paralinguistic vocalizations may inform later development with the potential to be a useful component to early detection of ASD. For the data collected in the study, a research assistant at the Georgia Institute of Technology labeled the segments using the various categories outlined in Table **??**. The reasoning behind relabeling the segments is to ensure that there is ground truth for the paralinguistic events and to use a majority vote based on the outputs of three voice activity detectors (VAD).

Table 3 Labels used for the segments using the annotation tool developed at Georgia Institute of Technology for the IBIS dataset.

Type	**Category of sound event**
Child	Speech, other vocalizations, fussing/crying, crying, laughter, other child
Adult	Male and female (near and far)
Noise	Toys, overlap, other

The importance of this dataset lies in the fact that the recordings were collected "in-the-wild" and constitute an important move forward in the scheme of validating models trained in laboratory environments, which are often sound-treated.

The MMDB dataset, which consists of speech, laughter, and crying samples, was used as the training data and the other two datasets were used as testing data. Table **??** shows the number of samples along with the durations (mean ± standard deviation) for all the datasets.

Table 4 Number of training and testing examples of MMDB, Strange Situation, and IBIS datasets for speech, laughter, and fussing/crying along with the mean and standard deviation of duration of the samples.

Dataset	Type of vocalization	Number of samples (N)	Duration(mean±standard deviation)
MMDB	Speech	200	1.14±0.66
	Laughter	128	1.31±1.28
	Fussing/Crying	142	2.65±4.21
Strange Situation	Speech	171	1.23±0.92
	Laughter	11	1.12±0.90
	Fussing/Crying	129	1.68±0.83
IBIS	Speech	510	1.23±0.92
	Laughter	48	1.12±0.90
	Fussing/Crying	421	1.68±0.83

1.2 Long-term intensity-based feature

A new measure to capture the long-term periodic structure of laughter using the energy or intensity contour is introduced below. The work by [**?**] uses *a priori* information about the frequency range (4–6 Hz) in which the sonic structure of laughter is apparent in the magnitude spectrum of the intensity contour of laughter. The advantage of this measure is that it is not dependent on the bandwidth of the audio signal and can be generalized for signals recorded at various sampling rates. The *apriori* information about the frequency with which the sonic structure manifests will not be used but uses window lengths of varying sizes that can encompass different syllable lengths. In the first step, the intensity or energy contour of the speech signal is computed using a Hamming window of 30 ms length and 10 ms overlap as shown in (**??**).

$$E[n] = \sum_{n=1}^{n} x[n]^2 \tag{1}$$

, where $x[n]$ is the windowed speech signal frame and $E[n]$ is the energy or intensity of the signal.

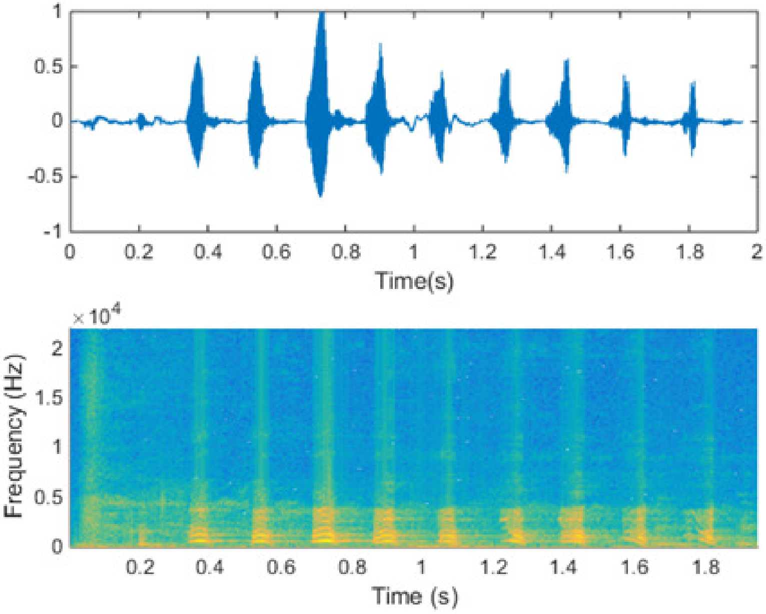


Fig. 4 Waveform of laughter sample from the MAHNOB [**?**] database along with the spectrogram displayed below it.

In Figure **??**, the repetitive structure of laughter can clearly be seen in the spectrogram, while such a structure was not apparent for speech as seen in Figure **??**. Using the intensity contour, the Hamming window length was again varied from 5 to 45 frames (in steps of 4) for children's laughter with different overlap window lengths. The reason for using different window lengths is due to the fact that these were the ranges of window lengths that resulted in good accuracies as will be discussed in Section **??**. From this syllable-level segment, the autocorrelation of the

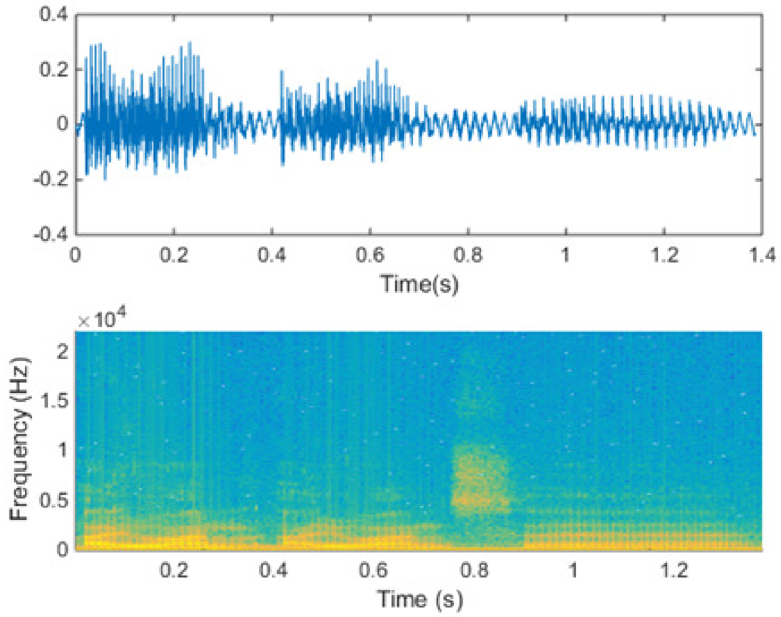


Fig. 5 Waveform of speech sample from the MAHNOB database [**?**] along with the spectrogram displayed below it.

intensity contour is computed as shown in (**??**).

$$R_{xx}[j] = \sum_n x_n \bar{x}_{n-j} \tag{2}$$

Then, a polynomial regression curve was fitted to the one-sided autocorrelation function and the absolute error was computed between the curve and the autocorrelation function. The idea behind computing the error was that the greater the periodic structure of the signal, which would be the case for laughter, the higher would be the error than for speech. Since the children's audio signals might consist of noise or cross-talk, we varied the degree, d, of the polynomial regression curve from 1 to 3. Also, for the children's speech there were four different overlap window lengths used ranging from 12.5% to 50% overlap. This resulted in 36 low-level descriptors for children's speech. There were 14 statistical measures computed from the features and these are shown in Table **??**.

Table 5 Statistical measures evaluated for syllable-level intensity features.

Statistical Measure
Arithmetic mean, median, mode, standard deviation, maximum and minimum values, flatness, skewness, kurtosis, 25^{th} quartile, 75^{th} quartile, inter-quartile ranges, 1^{st} percentile, 99^{th} percentile

The baseline acoustic features were extracted using the open-source audio feature extraction tool, openSMILE [**?**]. There were 57 low-level descriptors (LLD), shown in Table **??** extracted using a 30 ms Hamming window with 10 ms overlap. The delta and delta-delta measure for each LLD was also computed and the number of LLDs was 171. There were 39 statistical measures, shown in Table **??**, computed from the LLDs for each sample. The dimensionality of the feature set using openSMILE was 6669.

Table 6 Spectral and prosodic acoustic features extracted using openSMILE.

Feature	Number of low-level descriptors
Log-energy	3
Magnitude of Mel-Spectrum	78
Mel-frequency Cepstral Coefficients	39
Pitch	3
Pitch envelope	3
Probability of voicing	3
Magnitude in frequency band ($0-250Hz$, $250-650Hz$, $0-650Hz$, $1000-4000Hz$, and $3010-9123Hz$)	16
Spectral Rolloff (25^{th} ,50^{th}, 75^{th}, and 90^{th} percentile)	12
Spectral Flux	3
Spectral Position (Centroid, Maximum, and Minimum)	3
Zero-Crossing Rate	3

Table 7 Statistical measures evaluated for openSMILE features.

Statistical Measure
Max./Min. value and respective relative position within input, range, arithmetic mean,3 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, centroid, variance, number of non-zero elements, quadratic, geometric, absolute mean, arithmetic mean of contour and non-zero elements of contour, 95^{th} and 98^{th} percentiles, number of peaks, mean distance from peak, mean peak amplitude, quartile 1 - 3, and 3 inter-quartile ranges.

1.3 Results

Models were trained using the MMDB dataset and tested the models on the Strange Situation and IBIS datasets. The results will be discussed in two categories; the first set of results deals with classifying laughter against combinations of various categories (speech, fussing/crying, and non-laughter which consists of speech and fussing/crying) using only the top 50 features ranked by CFS syllable-level intensity features and the second category will be the combination of baseline acoustic and syllable-level features by ranking the top 100 features using CFS. The selected features for the three classification tasks are shown in Figure **??**.

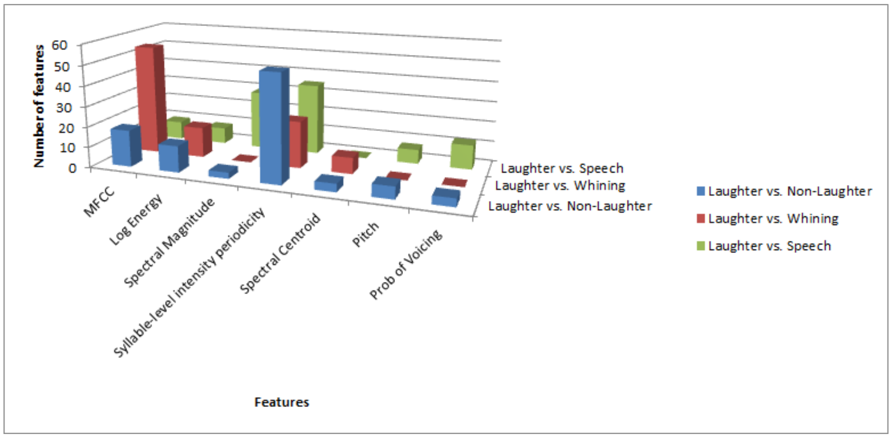


Fig. 6 Features selected for the three classification tasks viz. speech vs. laughter, fussing/crying vs. laughter, and non-laughter vs. laughter

Using the MMDB corpora for training, the results of the 10-fold cross validation are shown in Table **??** for the various classification tasks using the top 50 syllable-level features using CFS.

Table 8 Accuracy and recall of 10-fold cross-validation with training on MMDB corpus using the top 50 syllable-level features using a cost-sensitive linear kernel SVM classifier.

	Speech vs. Laughter	Whining vs. Laughter	Non-Laughter vs. Laughter
Accuracy	73.17%	71.85%	75.53%
Recall	72.23%	71.81%	74.63%

Using the MMDB corpora for training, the results of the 10-fold cross validation are shown in Table **??** for the various classification tasks using the top 50 syllable-level features using CFS.

Table 9 Accuracy and recall of 10-fold cross-validation with training on MMDB corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier.

	Speech vs. Laughter	Whining vs. Laughter	Non-Laughter vs. Laughter
Accuracy	84.75%	79.25%	81.27%
Recall	84.82%	78.77%	80.04%

Using the MMDB corpora for training and testing on the IBIS , the results of the test sets are shown in Table **??** for the various classification tasks using the top 100 baseline and syllable-level features using CFS.

Table 10 Accuracy and recall of training on MMDB corpus and testing on IBIS corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier.

	Speech vs. Laughter	Whining vs. Laughter	Non-Laughter vs. Laughter
Accuracy	85.12%	81.02%	82.53%
Recall	85.26%	81.12%	79.94%

Using the MMDB corpora for training and testing on the Strange Situation corpus , the results of the test sets are shown in Table **??** for the various classification tasks using the top 100 baseline and syllable-level features using CFS.

Table 11 Accuracy and recall of training on MMDB corpus and testing on Strange Situation corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier.

	Speech vs. Laughter	Whining vs. Laughter	Non-Laughter vs. Laughter
Accuracy	84.06%	90%	83.6%
Recall	87.26%	90.41%	87.12%

The results indicate that the syllable-level features are capable of detecting laughter from speech, fussing/crying, and , when both these events are treated as a single class, non-laughter to a reasonably high degree of accuracy and more importantly, a high recall rate as well. The significance of these results lie in the fact that the features trained on the MMDB dataset generalize well when applied to the Strange Situation and IBIS datasets which consists of data recorded in completely different conditions, subjects with a different age group, and with subjects at risk of ASD.

2 Multi-modal Laughter Detection in Toddlers' Speech When Interacting With Caregivers

2.1 Introduction

Smiling is one of the most common facial expressions used while interacting with friends or peers [?]. Smiles can manifest as Duchenne smiles, activated using the *Zygomaticus Major* and *Orbicularis Oculii* muscles concurrently, which are used to express positive affect. When only the *Zygomaticus Major* muscle is activated, the smile is considered to be forced [?]. Smiles, like laughter, can also be used to mask the true affective state of an individual. False smiles can be used to indicate that a person is happy while masking the true affective state which could range from deception to disgust [?].

There is limited understanding about the interaction between smile and laughter and one [?] hypothesis is that smiles have their origins in the silent bared-teeth submissive grimace of primates, and laughter evolved from the relaxed open-mouth display. Since, spontaneous smiles have been linked with laughter [?], an attempt has been made to use the information about smiles to reduce false positives in detecting laughter using only the audio modality.

The research by [?] discusses about performing multi-modal laughter detection in adults' speech and shows the improvement obtained from fusing the features from the audio and vision modalities compared to using either one of them. A logical extension of this work would be to analyze the data from children's interactions with caregivers. Previous research on smiling type and play type during parent-infant play has shown varying conclusions about the frequency of smiling with infants smiling more at the mother compared to the father during visual games, object play, and social games. While research which showed smiling preference for fathers involved games of physical and idiosyncratic nature.

2.2 Database

The MMDB corpus was used for the purpose of analysis and the modalities used were the audio from the lavalier microphones and the Canon side-view cameras for analyzing the smiles of the child. For the purposes of detecting laughter, the problem was treated as a laughter vs. non-laughter classification problem where the non-laughter elements included child's speech and fussing/crying. There were a number of difficulties experienced while analyzing the videos of the child. One major problem was that OMRON's smile tracker was used to initialize the face of the child automatically and given that the parent was also in the view of the camera, the parent's face would be mistaken for the child's face. To overcome this issue, a manual selection of the child's face was done by selecting the frame when the child's face was detected by the smile tracker. This process mitigated the false positives of the child's face being detected. The other issues that were faced while detecting the child's face were when the face was obscured from the view of the camera due to the examiner or parent moving in front of the child, the child turning his or her face away from the view of the camera, or the child moving away from the view of the camera by getting distracted by an object in the room. These were issues that could potentially be addressed by using information from the AXIS cameras, but that would be pertinent to whether the child's face can be accurately detected using them.

Having detected the child's face and extracting the information about the smile, the child's speech annotations were lined up with the frame-level results of the Canon videos. The annotations in ELAN are relative to the Canon videos and therefore the synchronization is a simple process of lining up the various events belonging to other modalities. Once the annotations have been lined up, we need to take into account that the smile detector can produce false negatives due to the tracker failing to track the face when the child's face is in view. For this purpose, we used a threshold method wherein only the laughter and non-laughter annotations are used when for more than 70% of the duration of the event, the smile detector produces a valid output (a vector of non-zero features).

2.3 Feature Extraction and Selection

The openSMILE features along with the syllable-level intensity features, described in Section **??**, were extracted from the laughter and non-laughter samples. For the visual features, the OMRON Okao smile detection system was used to extract the frame-level features and the feature that were used for analyses was the smile strength. There were two methods employed for feature selection. The first technique is the combination of the filter and wrapper-based techniques with the filter-based technique used being the correlation-based feature selection technique followed by the wrapper-based technique which is the sequential-forward selection method with a linear kernel SVM as the base classifier. The other technique employed was using a restricted Boltzmann machine (RBM) with contrastive divergence and this is widely used in image classification and of late, in speech recognition for the purposes of learning deep learning models.

An RBM is a undirected graphical model which consists of bipartite graphs. There are two types of variables in the architecture, a set of visible units, V, and followed by hidden units, H. There are no connections within V and H, as shown in Figure **??**, and thus each set of units is conditionally independent of the other.

For every possible connection between the binary visible, v, and hidden units, h, the RBM assigns an energy and this is given using the equation shown in (**??**)

$$E(v,h) = -\sum_{i,j} W_{ij} v_i h_j - \sum_i a_i v_i - \sum_j b_j h_j. \tag{3}$$

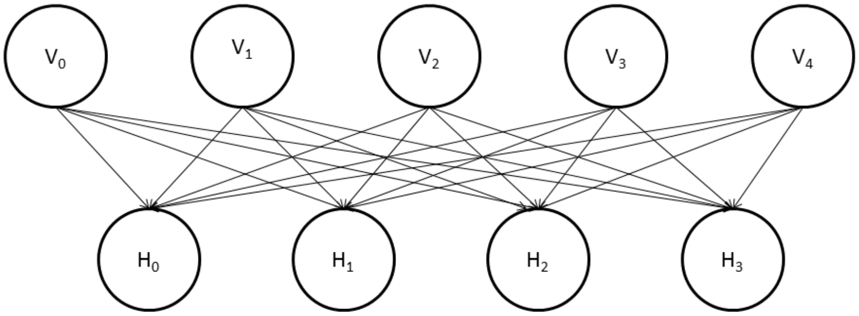


Fig. 7 Structure of a restricted Boltzmann machine (RBM) with connections between visible layer, V, and hidden layer, H.

where v_i and h_j are the binary states of the visible unit i and hidden unit j. The a and b are the biases of the visible and hidden units respectively. W_{ij} represents the weights or the strength between the visible and hidden units.

The conditional probabilities of each of the visible and hidden units is given in (**??**) and (**??**),

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum\nolimits_i W_{ij} v_i) \tag{4}$$

$$p(v_i = 1 \mid h) = \sigma(a_i + \sum\nolimits_j W_{ij} h_j) \tag{5}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

is the logistic function.

The probability that is assigned to every possible joint configuration (v,h) is given in (**??**),

$$p(v,h) = \frac{e^{-E(v,h)}}{Z} = \frac{e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}} \tag{7}$$

where Z is the partition function. The marginal distribution of the visible units is given as

$$p(v) = \sum\nolimits_h p(v,h) \tag{8}$$

and the gradient of the average log-likelihood is given as

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_\infty \tag{9}$$

The $< . >_\infty$ cannot be computed efficiently as it involves the normalization constant Z and it is a sum of over all configurations of the variables making the problem intractable. This can be avoided by using the contrastive divergence (CD) algorithm by sampling from the distribution using Gibbs sampling. This involves setting the initial values of the visible units to the feature set and then sampling the hidden units given the visible units. After this, the visible units are then sampled using the hidden units and the process is alternated between the two. This is shown in Figure **??**.This sampling requires using the conditional distributions given in (**??**) and (**??**) which are easy to compute. The CD algorithm is given as,

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_k \tag{10}$$

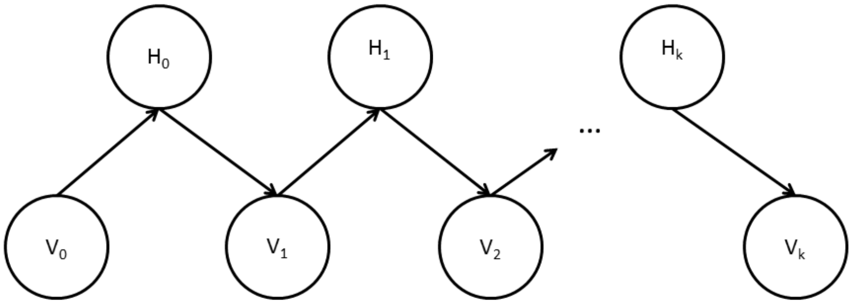


Fig. 8 Working of the contrastive divergence (CD) algorithm between the hidden and visible units in an RBM.

For the purposes of research in this section, the Gaussian- Bernoulli RBM was used to deal with feature sets that used acoustic and visual modalities. In this method, the visible units are treated as originating from a Gaussian distribution and the hidden units are binary. The equation of the energy function becomes,

$$E(v,h) = -\sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} h_j W_{ij} - \sum_j b_j h_j. \tag{11}$$

The conditional probabilities of the visible and hidden units are modified as shown in (**??**) and (**??**).

$$p(v_i = v \mid h) = \mathscr{N}(v \mid a_i + \sum_j W_{ij} h_j, \sigma_i^2) \tag{12}$$

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum_i W_{ij} \frac{v_i}{\sigma_i^2}) \tag{13}$$

where $\mathscr{N}(\cdot \mid \mu, \sigma^2)$ is a Gaussian probability density function with mean μ and variance σ^2.

2.4 Methodology

Two feature selection methodologies for the multi-modal analysis were employed. In the first part, as shown in Figure **??**, we used the CFS on the acoustic features and concatenated with the visual features followed by passing the feature set through a sequential forward selection (SFS) with the base classifier being a linear kernel SVM.

The features selected using this scheme is shown in Table **??** and include spectral centroid, syllable-level intensity, and smile confidence features.

Table 12 Acoustic and visual features selected using feature selection based on combination of filter and wrapper-based methods using the MMDB dataset.

Feature	Number of features selected
Spectral centroid	2
Syllable-level Intensity Autocorrelation Error	1
Smile confidence	1

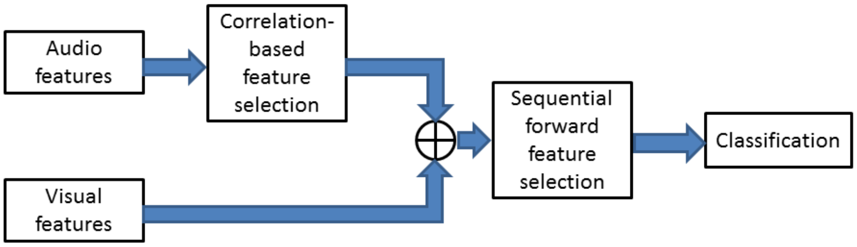

Audio features
Correlation-based feature selection
Visual features
Sequential forward feature selection
Classification

Fig. 9 Architecture of the system employed for multi-modal laughter detection using combination of filter and wrapper-based feature selection schemes.

For the multi-modal analysis using RBMs, the method employed is the bimodal deep belief network (DBN) architecture [**?**]. Here, the lower layers learn the audio and video features separately followed by concatenating and feeding them to another RBM, as shown in Figure **??**, which learns the correlations between the various modalities. For this architecture, we employed the Gaussian-Bernoulli RBM for the first layers followed by a Bernoulli-Bernoulli RBM for the top-most layer. This is a similar architecture that has been previously used in multi-modal emotion recognition by [**?**]. The only parameter being varied is the number of hidden units with all the other parameters such as learning rate, number of iterations for the CD algorithm, and batch size being constant. The number of hidden units varied from 10 to 50 with a step size of 10. A grid search is performed for finding the configuration of the number of hidden units for each RBM that results in the best accuracy using a 10-fold cross-validation scheme.

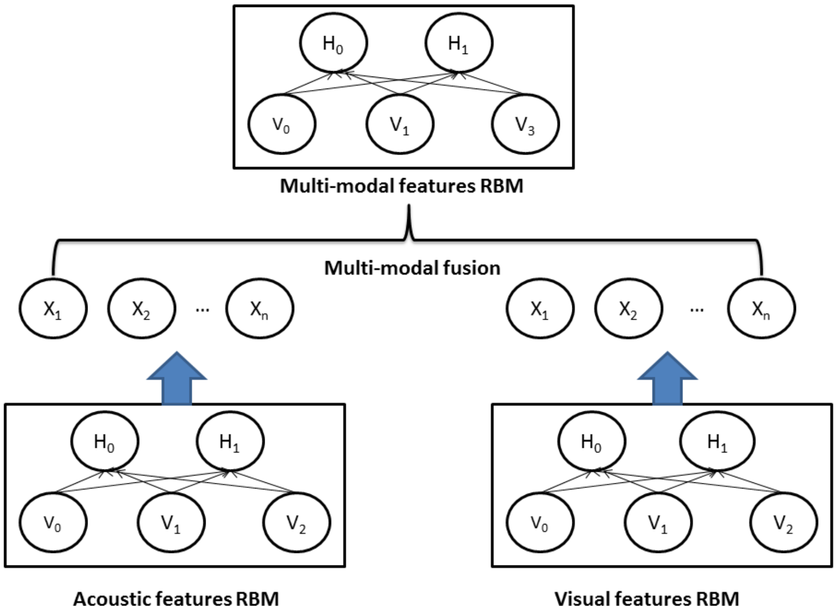


Fig. 10 Architecture of the system employed for multi-modal laughter detection using RBMs.

2.5 Results

Owing to the fact that the number of samples used in this study was small due to the various limitations in analyzing the videos as described earlier, a 10-fold cross-validation was performed on the dataset with a linear kernel SVM. Considering the imbalance in the training data, we used a cost-sensitive classification scheme with the cost matrix given as,

$$C = \begin{bmatrix} 0 & 1 \\ 1.81 & 0 \end{bmatrix} \tag{14}$$

Classification using the acoustic features from the filter based method, where the top 100 audio features are ranked, resulted in a confusion matrix for laughter vs. non-laughter as shown in Table **??**.

Table 13 Accuracy and Recall of the 10-fold cross validation results using SVM for the audio, video, and audio-video modalities.

Modality	Accuracy	Recall
Audio	78.8%	77.14%
Video	81.1%	81.85%
Audio + Video	86.17%	85.48%

The accuracy is **86.2%** which this is significantly higher than using the features from either modality alone. The recall rate for the non-laughter class is significantly higher than either of the two modalities but the one for laughter is slightly lower than that of visual modality alone. Nonetheless, these results are indicative that the use of multi-modal information would definitely enhance the classification over using either of the modalities alone.

The best results were obtained using 40 hidden units for the speech RBM, 10 hidden units for the visual features RBM, and finally 25 hidden units for the top most RBM which uses the outputs of the speech and visual RBMs. The outputs of the RBMs are then fed as features to an SVM classifier. The results are shown in Table **??**.

Table 14 Accuracy and Recall of the 10-fold cross validation results using RBMs and SVM classifier for the audio, video, and audio-video modalities.

Modality	Accuracy	Recall
Audio	83.41%	81.88%
Video	80.18%	9.96%
Audio + Video	88.94%	87.62%

With the use of the RBM architecture, the accuracy of the system is **88.94%** and the recall rate for non-laughter, **92.14%**, is better than that of the previous methodology.

The research has focused on using multi-modal information for the detection of laughter in children's speech while interacting with their caregivers in a semi-structured environment. The integration of visual features using the OMRON Okao smile tracking system has the ability to capture the smile characteristics in children's laughter. The audio and the vision modalities on their own are capable of discriminating between laughter from non-laughter events but when the features are combined, there is an improvement in the classification accuracy. The use of the multi-modal architecture using a restricted Boltzmann machine yields in a significant improvement in the accuracy over using an RBM for features of only one modality.

References

1. Ainsworth, M., Blehar, M., Waters, E., Wall, S.: Patterns of attachment. hills-dale. NJ Eribaum (1978)
2. Apple, W., Streeter, L.A., Krauss, R.M.: Effects of pitch and speech rate on personal attributions. Journal of Personality and Social Psychology **37**(5), 715 (1979)
3. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. The Journal of the Acoustical Society of America **110**(3), 1581–1597 (2001)
4. Darwin, C.: The expression of the emotions in man and animals. Oxford University Press (2002)
5. Esposito, G., Venuti, P.: Comparative analysis of crying in children with autism, developmental delays, and typical development. Focus on Autism and Other Developmental Disabilities **24**(4), 240–247 (2009)
6. Estes, A., Zwaigenbaum, L., Gu, H., John, T.S., Paterson, S., Elison, J.T., Hazlett, H., Botteron, K., Dager, S.R., Schultz, R.T., et al.: Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. Journal of neurodevelopmental disorders **7**(1), 1 (2015)
7. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia, pp. 1459–1462. ACM (2010)
8. Hess, U., Bourgeois, P.: You smile–i smile: Emotion expression in social interaction. Biological psychology **84**(3), 514–520 (2010)
9. Hudenko, W.J., Stone, W., Bachorowski, J.A.: Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder. Journal of Autism and Developmental Disorders **39**(10), 1392–1400 (2009)
10. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3687–3691. IEEE (2013)
11. Kraut, R.E., Johnston, R.E.: Social and emotional messages of smiling: An ethological approach. Journal of personality and social psychology **37**(9), 1539 (1979)
12. Lockard, J., Fahrenbruch, C., Smith, J., Morgan, C.: Smiling and laughter: Different phyletic origins? Bulletin of the Psychonomic Society **10**(3), 183–186 (1977)
13. Meadows, C.: Psychological Experiences of Joy and Emotional Fulfillment. Routledge (2013)
14. Mehu, M., Dunbar, R.I.: Relationship between smiling and laughter in humans (homo sapiens): Testing the power asymmetry hypothesis. Folia Primatologica **79**(5), 269–280 (2008)
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696 (2011)
16. Oh, J., Cho, E., Slaney, M.: Characteristic contours of syllabic-level units in laughter. In: INTERSPEECH, pp. 158–162 (2013)
17. Oller, D.K., Niyogi, P., Gray, S., Richards, J.A., Gilkerson, J., Xu, D., Yapanel, U., Warren, S.F.: Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences **107**(30), 13,354–13,359 (2010)
18. Orozco, J., García, C.A.R.: Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. In: European Symposium on Artificial Neural Networks, Bruges (Belgium), pp. 349–354 (2003)
19. Petridis, S., Martinez, B., Pantic, M.: The mahnob laughter database. Image and Vision Computing **31**(2), 186–202 (2013)
20. Poyatos, F.: Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds, vol. 92. John Benjamins Publishing (1993)
21. Prince E.B., C.A.G.D.M.K.R.A.R.J.R.J., Messinger, D.: Automated measurement of dyadic interaction predicts expert ratings of attachment in the strange situation. Association for Psychological Science Annual Convention (2015)

22. Ranganath, R., Jurafsky, D., McFarland, D.A.: Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Computer Speech & Language **27**(1), 89–115 (2013)
23. Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Scalaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C.H., Rao, H., Kim, J., Presti, L., Zhang, J., Lantsman, D., , Bidwell, J., Ye, Z.: Decoding children's social behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. IEEE (2013)
24. Rothbart, M.K.: Laughter in young children. Psychological bulletin **80**(3), 247 (1973)
25. Szameitat, D.P., Alter, K., Szameitat, A.J., Wildgruber, D., Sterr, A., Darwin, C.J.: Acoustic profiles of distinct emotional expressions in laughter. The Journal of the Acoustical Society of America **126**(1), 354–366 (2009)
26. Tepperman, J., Traum, D., Narayanan, S.: " yeah right": Sarcasm recognition for spoken dialogue systems. In: Ninth International Conference on Spoken Language Processing (2006)
27. Waters, E.: The reliability and stability of individual differences in infant-mother attachment. Child Development pp. 483–494 (1978)
28. Wolff, J.J., Gu, H., Gerig, G., Elison, J.T., Styner, M., Gouttard, S., Botteron, K.N., Dager, S.R., Dawson, G., Estes, A.M., et al.: Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. American Journal of Psychiatry (2012)