PARALINGUISTIC EVENT DETECTION IN CHILDREN'S SPEECH

A Thesis
Presented to
The Academic Faculty

by

Hrishikesh Rao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2015

PARALINGUISTIC EVENT DETECTION IN CHILDREN'S SPEECH

Approved by:

Dr. Mark A. Clements, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Elliot Moore II
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Irfan Essa
School of Interactive Computing
Georgia Institute of Technology

Dr. David Anderson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Agata Rozga
School of Interactive Computing
Georgia Institute of Technology

Date Approved: 05 October 2015

To my parents and Jyotsna

ACKNOWLEDGEMENTS

This thesis would not have been possible without the contributions of some wonderful people that I have worked with and met in my life at Georgia Tech. I would like to thank Dr. Mark Clements for being my advisor and mentor during my Ph.D. This thesis would not have been possible without his contributions in formulating the thesis topic and his pearls of wisdom that I was fortunate enough to obtain during our Tuesday morning meetings. I couldn't have asked for a better advisor than him.

I would like to thank Dr. David Anderson who referred me to Dr. Clements and also someone that I had the opportunity to work with in Summer 2010. I will be forever indebted to him for all that he has done to help me. I would also like to thank Dr. Elliot Moore II and Dr. Irfan Essa for being on my committee and their immensely useful feedback provided to me during the course of my research.

I would also like to thank Agata Rozga for guiding me throughout the journey with her feedback about my research and for her invaluable contribution in acquiring the various datasets that have been used in the analyses. A special thank you to the wonderful collaborators (James Rehg, Gregory Abowd, Yin Li, Zhefan Ye, Chanel Bridges, and Audrey Southerland) in the Expeditions project for their gracious help and support during the project.

My time at Georgia Tech would not be complete without the company of many people that I have befriended over the years which helped me to keep distracted from the ups and downs of a graduate student's life.

Finally, I would like to thank my parents and my fiancee, Jyotsna, for all their sacrifice, love, and support during times of trials and tribulations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Paralinguistic events are useful indicators of the affective state of a speaker. These cues, in children's speech, are used to form social bonds with their caregivers. They have also been found to be useful in the very early detection of developmental disorders such as autism spectrum disorder (ASD) in children's speech. Prior work on children's speech has focused on the use of a limited number of subjects which don't have sufficient diversity in the type of vocalizations that are produced. Also, the features that are necessary to understand the production of paralinguistic events is not fully understood. To account for the lack of an off-the-shelf solution to detect instances of laughter and crying in children's speech, the focus of the thesis is to investigate and develop signal processing algorithms to extract acoustic features and use machine learning algorithms on various corpora. Results obtained using baseline spectral and prosodic features indicate the ability of the combination of spectral, prosodic, and dysphonation-related features that are needed to detect laughter and whining in toddlers' speech with different age groups and recording environments. The use of long-term features were found to be useful to capture the periodic properties of laughter in adults' and children's speech and detected instances of laughter to a high degree of accuracy. Finally, the thesis focuses on the use of multi-modal information using acoustic features and computer vision-based smile-related features to detect instances of laughter and to reduce the instances of false positives in adults' and children's speech. The fusion of the features resulted in an improvement of the accuracy and recall rates than when using either of the two modalities on their own.

CHAPTER I

INTRODUCTION

Researchers in voice recognition and understanding have focused mainly on speech recognition systems that identify“what” is being said. Recently, a plethora of work has been focused on trying to detect emotion or affect in speech to help understand “how” speech is uttered under various conditions. Emotion recognition from speech can be potentially useful for both diagnostic and commercial applications. Depression, a major cause of disability and loss of productivity in adults [3], affects the pitch, speaking rate, loudness, and various articulation gestures in adult speech. Conversations involving customer service representatives and clients or consumers can be monitored for the affective state of the latter to gauge the performance of representatives in resolving issues related to service. Of late, there has been a focus on analyzing the non-verbal aspects, or paralinguistic cues, of human speech for affective classification.

Paralinguistic cues are the non-phonemic aspects of human speech that can be used for changing the semantic content of an utterance. These cues are characterized by such signatures as modulation of pitch, amplitude, and temporal patterns in speech. Humans are capable of interpreting affective information from speech but, are likely also to rely on the paralinguistic component in speech [4]. The paralinguistic cues in human speech encompass a wide variety of differentiators that are discussed in [5]. These cues, shown in Table 1, characterize physiological and emotional states and are produced naturally or voluntarily.

These paralinguistic phenomena can be considered to have varied nuances in the way they are produced and used as a result of the emotional state of the speaker or

Table 1: List of paralinguistic differentiators in human speech.

Paralinguistic Differentiators
Laughter, crying, shouting, sighing, gasping, panting, yawning, coughing, spitting, belching, hiccuping, and sneezing.

to change the semantic content of an utterance. For instance, laughter, as described in Charles Darwin's seminal work on emotions [6], is primarily used to express joy or happiness. This may occur as a result of a visual stimulus (watching a situational comedy or "sitcom" on television), auditory stimulus (listening to a stand-up comedian in a nightclub), or physical stimulus (tickling). Thus, laughter can be thought of as being related to social bonding. On the other hand, laughter can also be used to be dismissive or to ridicule someone, which is often seen in television debates and can also be used in cases of "schadenfreude" or "malicious pleasure" where one may use laughter to express joy at someone else's misfortune. A spoken phrase with laughter included would be interpreted by the listener as a phrase spoken in jest or, if modulated with whining, it could be perceived as frustration in speech. The phrase, "Yeah right" is an example of a commonly used utterance for expressing sarcasm. It has a positive literal meaning, though the semantic value is negative. The study by Tepperman et al. 2006 [7] found that laughter was an important contextual feature in identifying sarcasm in speech and the phrase, when used in conjunction with laughter or in adjacent turns of laughter with either speaker, was an important objective cue for detecting sarcasm. When it comes to visual cues, research has shown that facial expressions conveying emotions are innate rather than being acquired during the growing process. This has led to the development of the theory of the universality of emotions with various facial expressions representing the same type of emotion being expressed across various cultures [8]. A fine-grained distinction within a particular facial expression can be useful in characterizing social behavior. For instance, smiling is generally thought of as expressing joy or happiness, but can also be used to avoid an awkward situation. Research by Ekman et al. 1990 [9] has supported the theory of distinguishing between various types of smiling rather than treating a facial expression such as smiling as being representative of a single class of behavior.

The main focus of this work is to detect laughter and fussing/crying in toddlers' speech using acoustic features, and to explore the use of long-term acoustic features that characterize the periodic structure of laughter. The work also focuses on using multi-modal information to detect laughter in children's as well as adults' speech to detect laughter using acoustic and computer vision-based features. 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. It 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 [10]. Fussing/Crying could indicate that the child is upset or disinterested in the task being initiated by the caregiver in a dyadic setting. These cues have been found to be important markers in the very early detection of autism spectrum disorder (ASD) [11, 12], and the diarization of such events in extended recordings can be a useful aid in the diagnosis of developmental disorders [13, 14].

CHAPTER II

BACKGROUND AND PRIOR WORK

2.1 Infant Vocal Development

The advent of vocal development takes place in the infant's first year of life. The models of infant vocal development were proposed in the 1970s and 1980s by many researchers [1, 15, 16, 17, 18]. These models include similar vocalization types, ages of emergence, and a number of levels of vocal development that are widely accepted benchmarks, but that have operational differences and distinct terminologies. The evolutionary path for infant vocal development, shown in Figure 1, can be described using the four-stage model prescribed by Oller et al. 1999 [1]. This model does not take into account vegetative sounds such as coughing, sneezing, and burping, or fixed vocal signals such as crying and laughter. The vocalizations which are considered as precursors to speech produced by the infants are described in the model as protophones. It is of interest to note that vegetative sounds and fixed vocal signals are present in other species, while protophones are unique to humans.

Research by Nathani et al. 2006 [2] has shown that inclusion of non-speech vocalizations in the model of language development provides foundational elements used for speech production. The Stark Assessment of Early Vocal Development-Revised (SAEVD-R) scale uses five levels to differentiate the different landmarks in infant vocal development. The scale, in which the progression of protophones and non-speech vocalizations is described, is shown in Figure 2.

The progression of speech begins with Level 1, the production of quasi-resonant (QR) nuclei that are faint, low-pitched grunt-like sounds cannot be transcribed as vowels. The QR nuclei are characterized by the lack of energy above 2000 Hz. In

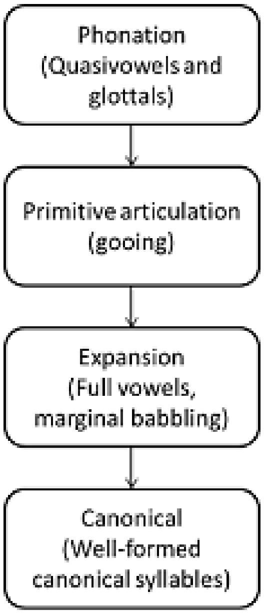
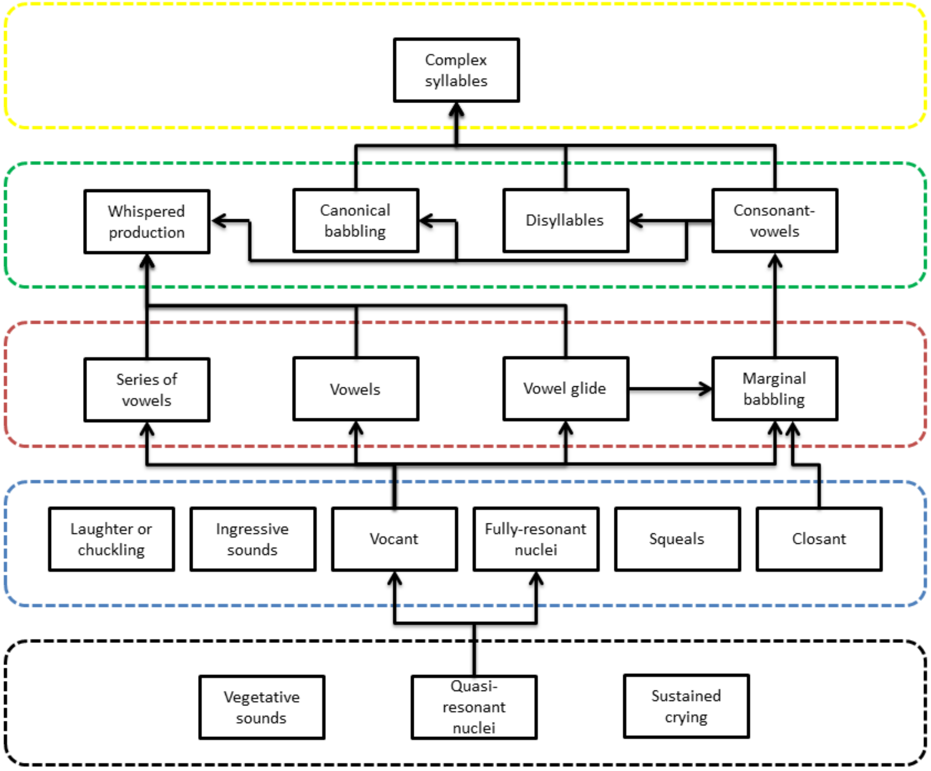

Phonation
(Quasivowels and glottals)
Primitive articulation
(gooing)
Expansion
(Full vowels, marginal babbling)
Canonical
(Well-formed canonical syllables)

Figure 1: Stages of infant vocal development proposed by Oller et al. 1999 [1]

Level 2, at 1-4 months of age, infants develop control over the production of sounds using their vocal tract. Protophones in this stage are of a longer duration than QR nuclei and are fully-resonant (FR) with energy over a wide range of frequencies. The third level, the expansion phase, which occurs at 3-8 months of age, consists of isolated vowels that are longer than QR and FR and are fully transcribable. This stage also marks the beginning of marginal babbling. The fourth level, occurring at 5-10 months of age, consists of babbling that is a repetitive production of consonants and vowels with adult-like formant transitions. The fifth level consists of the production of syllables with complex articulatory and phonatory characteristics and occurs just before production of words.

The majority of the paralinguistic cues are produced during the first stage and these include vegetative sounds such as laughter, coughs, and sneezes. Sustained crying and fussing occur in the first two months of life (Level 1) and are produced when the child is hungry or in pain. Laughter and chuckles are produced during the stage of controlled phonation (Level 2) as the child develops voluntary control of the vocal tract.



Type of vocalization	Description
Vegetative sounds	Burps, coughs,sneezes,etc.
Quasi-resonant nuclei	Faint, low-pitched grunt-like sounds with muffled resonance. Lack of energy >2000 Hz
Vocant	Vowel-like segment
Vowel glide	Vocant with slow formant transition (>200 ms)
Marginal babbling	Series of vocants and closants
Canonical babbling	Repeated consonant-vowel segments

Label	Level	Age (months)
⬛	Reflexive	0-2
🟥	Control of phonation	1-4
🟦	Expansion	3-8
🟩	Basic Canonical Phase	5-10
🟨	Advanced Forms	9-18

Figure 2: Stages of infant vocal development proposed by Nathani et al. 2006 [2]

2.2 Role of Paralanguage in Children's Speech

Affective expression has been viewed as serving a function for cognitive development and is suggested to be non-dissociable. In an infant's vocalizations, the term "functional flexibility" [19] is employed to characterize the affective nature of vocalizations. Functional flexibility in infants implies that the change in facial affect associated with an infant's vocalization should correspond to an observable change in the communicative act of the infant and the caregiver's action in response to the infant's social act. Research by Oller et al. 2013 [19] has shown that protophones such as squeals, growls, and vocants or vowel-like sounds were primarily rated as neutral with some cases of positive and negative affect.

Laughter, a rhythmic smile-linked vocalization, resulted in an overwhelming positive affect while crying, resulted in negative affect. In the first six months of life, infants normally produce laughter in response to intense auditory or tactile stimulation and in the second half of the first year; laughter is produced in response to subtle and complex social and visual stimulation. Crying is one of the vocal behaviors that promotes proximity with the mother and is considered as a part of attachment behavior [20]. Crying generally arouses alarm or displeasure in an infant and is used to elicit intervention to terminate its recurrence. Crying has been characterized as a sequence of inspiratory and expiratory phonation episodes with the former being short in duration while the latter is of a long duration involving phonation, dysphonation, or hyperphonation of a long duration [21].

2.3 Interaction of vocal and facial cues in children's paralanguage

Paralinguistic events can be thought of not only emanating from one source such as vocal, facial, or body movement cues but as a result of an interaction between all of them. Studies have found laughter to be a result of stereotyped exhalation of air outside the mouth cavity along with rhythmic head and body movements [22]. Laughter can also be thought of being linked with smiling even though they have different phylogenetic origins [23]. Research by Scarpa et al. 1997 [24] has shown differences in heart rate and skin conductance in three-year old children who exhibit inhibitory behaviors such as crying compared to no crying behavior being displayed. From the point of view of multi-modal analysis, using the speech and vision modalities would be of significant use considering the plethora of work being done in extracting acoustic and visual features and applying these techniques to children's social behaviors. The research in this thesis focuses on using these two modalities to detect laughter.

Smiling has been hypothesized to have evolved from the silent bared-teeth display of chimpanzees and laughter was likely to have evolved from the relaxed open-mouth display or play face shown by non-human primates during play encounters. The functions of smiling varies across various primates from being restricted to submission or appeasement [25, 26] to performing a socio-positive function. Laughter on the other hand is a function of social play [27, 28]. In infants, different types of smiles have been found to be used for different play types between parents and infants [29]. Work done by Messinger et al. 1999 [30] has shown that infants produce more Duchenne smiles (contraction of zygomatic major and orbicularis oculi muscles) than non-Duchenne ones and found that more than half of the Duchenne smiles involved opening of the mouth which could be a precursor to laughter. Smiling and laughter have been shown [31] to be produced by toddlers in the presence of other children and adults and serves to form social bonds. The major difference that was noted in the research was that smiling occurred as an accompaniment to incidental events than was the case with laughter which was produced in response to events that were deemed to be frivolous.

2.4 Databases

A considerable amount of research has been devoted to the study of adults' paralinguistic cues and their role in detecting the affective state of the speaker. Recent research has focused on detecting laughter in various corpora such as ICSI [32], AMI [33], AVLC [34], and MAHNOB [35]. These databases consist of recording from multi-participant meetings using a single (audio) or multiple (audio and video) modalities. Also, these databases have recordings which are of spontaneous or simulated in nature.

Databases involving children's speech with a large sample size are scarce. The main challenge involving the analyses of data of children's speech is the variation in the vocalizations during a child's development in the early stages. This may result in having an acoustic feature space that may not generalize well to data from children at a different stage of development. Since the focus of this research is to detect paralinguistic cues such as laughter and crying in children's speech, a description of the various corpora in which these cues are analyzed would be of interest. An early work by [36] analyzed the laughter produced by children at three years of age. The study involved analyzing children during their interaction with the mother for two episodes of 30 minutes each. The study was an attempt at studying the acoustic characteristics of various types of laughter; comment, chuckle, rhythmical, and squeal.

The number of the subjects in the study was low (*N=4*). Subsequently, research by Hudenko et al. 2009 [11] involved analyzing the differences in laughter in children with ASD (*N=15* and 8 to 10 years of age) with those of typically developing children. Furthermore, in [19] the pre-linguistic vocalizations from nine infants in a longitudinal study at different stages of their development (3 to 5, 6 to 7, and 10 to 12 months of age) were analyzed. The goal of the study was to analyze the emotional content of the vocalizations.

These databases involve analyses that are focused on a small set of subjects and are not automated in the detection of laughter. These issues were addressed in the work by Batliner et al. 2010 and Batliner et al. 2011 [37, 38], which had speech recordings from adolescents in a naturalistic setting wherein the subjects interacted with the Artificial Intelligence Robot (AIBO) by Sony. The robot was controlled by a human operator and was made to perform a fixed, pre-determined sequence of actions. Crying, on the other hand, has been analyzed for detection of developmental and pathological disorders such as hearing loss and hypothyroidism [39]. It has also been used for identification of infants [40]. The automated classification of crying and non-crying sounds in infants' speech was studied by [41] with acoustic analysis and machine learning techniques applied to recordings of children in a pre-school environment. Research by Abou-Abbas 2015 [21] used crying data from 1 to 53 day old infants recorded in hospitals in Canada and Lebanon and consisted of subjects who were healthy and those having a pathological condition.

Most of these databases consist of recordings in a laboratory environment which is noise-free or with a low number of subjects. Recordings from real-world environments, though desirable, are difficult to obtain due to privacy issues and the laborious nature of encoding vocalizations. The ideal middle ground would be to analyze databases, such as the Multimodal Dyadic Behavior Dataset and Strange Situation which are described in Sections 4.1.1 and 5.2 , which consist of children's speech and paralinguistic samples with varying degree of background noise and cross-talk to get a sense of the generalization properties of the acoustic features and the models developed using machine learning techniques.

2.5 Findings

Owing to the differences in the acoustic feature space due to the development in the child's articulatory and phonatory system, it is imperative to have an understanding of the acoustic features that would help characterize the child's speech from laughter and crying. The study by Nwokah et al. 1993 [36] found no differences in the number and duration of laughter events in children when compared with adults. The key difference was the in the fundamental frequency which was in the higher range of female laughter (400-500 Hz). Research by Hudenko et al. 2009 [11] analyzed the frequency of voiced and unvoiced laughter in children with ASD and compared it with typically developing children and found that children with ASD produced almost no unvoiced laughter than the controls whose laughs were 37-48% unvoiced.The work by Batliner et al. 2010 and Batliner et al. 2011 [37, 38] extracted 5967 spectral and prosodic acoustic features for their work in discriminating laughter from children's speech for complete turns and at the word level. The acoustic features were extracted using the open-source acoustic feature extraction tool, openSMILE. The relevant features were selected by using a leave-one-subject-out methodology by computing the Pearson correlation coefficient between the features and the classes. The base classifier used was a support vector machine (SVM) and the accuracy for detecting laughter in children's speech was 82%. The relevant features for this task were the zero-crossing rate, energy, pitch, mel-frequency cepstral coefficients (MFCC), and distribution of signal energy among spectral bands. These features characterize the reoccurring nature of laughter and are describing a pattern of repeating change between voiced and unvoiced segments and associated changes in speech spectra. Crying, like laughter, has been shown to have a higher fundamental frequency compared to babbling at different stages of an infant's development in the first year. The work done by Ruvolo et al. 2008 [41] used spatio-temporal box filter features extracted from sonograms of crying episodes. These features capture the beat, rhythm, and cadence of crying which has a highly rhythmic structure. Using the Tabu feature selection method [42] on 2,000,000 features and Gentle-boost [43], the area under the receiver operating characteristic (ROC) curve was 94.67% for four-second clips and 97% for eight-second clips. The performance degrades with the decrease in the length of the crying clip with the accuracy falling below chance level (50%) for clips smaller than 600 ms. The work done by Abou-Abbas et al. 2015 [21] used a 7-state Hidden Markov Model (HMM) to detect instances of inspiratory and expiratory periods of newborn infants crying using 50-ms window of mel-frequency cepstral coefficients (MFCC) resulted in an accuracy of 78% and for only the expiratory period resulted in an accuracy of 84%.

2.6 Summary

The automatic detection of paralinguistic events is a relatively nascent topic compared to adults' paralinguistic event detection, and it poses several challenges owing to the fact that the recordings consist of sample sizes that are low in number. As described by Schuller et al. 2013 [44], which is relevant to an adult's paralinguistic analysis, but could also be said of children's paralinguistics, the key challenges are the coupling of tasks, novel feature extraction and robustness. The studies described show the use of some of the basic acoustic features, but not the entire gamut of features that can be employed. Another interesting aspect of this area of research is to build feature selection schemes that can characterize the nature of paralinguistic events and also generalize well to other datasets with subjects of different age groups and recording environments. It would be beneficial to use information from other modalities such as smile detection in computer vision for improving the analysis of laughter detection due to the simultaneous occurrence of both events. Multi-modal analysis could also help in obtaining high-level information about when speech occurs with smiling, which could potentially provide information about the affective nature of vocalization.

CHAPTER III

DATABASES

3.1 Introduction

For the purposes of the research in this thesis, several of databases involving children's interactions with their caregivers were employed. As described in Section 2.4, databases involving children's speech do not have sufficient diversity in terms of number of subjects, the age range, and the environments in which the data is collected. Owing to the fact that acoustic features which have been designed for adults' speech may not necessarily generalize well when applied to children's speech, the features that are required to study paralinguistic event detection in children's speech is quite poorly understood. This thesis has made an attempt to analyze data recorded in laboratory and 'in-the-wild' environments to gain an understanding of which features are robust enough to detect laughter and crying in children's speech when trained on data in clean environments and tested on data collected in noisy conditions. The purpose of this chapter is to enlighten the reader about the potential challenges a researcher might encounter owing to the differences in the way paralinguistic events are produced, the context in which they are produced, the recording environments, and the age group of the subjects. Merely building models on one dataset might not be sufficient to validate the accuracy when tested on data recorded in noisy conditions. This chapter focuses on three datasets involving children's speech and they are the Multi-modal Dyadic Behavior Dataset (MMDB), Strange Situation, FAU-Aibo Emotion Corpus (AEC), the Weill Cornell Medical College (WCMC), the Oxford Vocalizations (OxVoc) Sounds and the Infant Brain Imaging Study (IBIS) datasets. Along with these datasets, the SSPNet Vocalizations Corpus (SVC) consisting of adults' laughter and fillers using only the audio modality and MAHNOB Laughter database which consists of multi-modal data recordings of adults' laughter was also used to validate the syllable-level acoustic features and multi-modal detection of laughter which will be discussed in the future chapters.

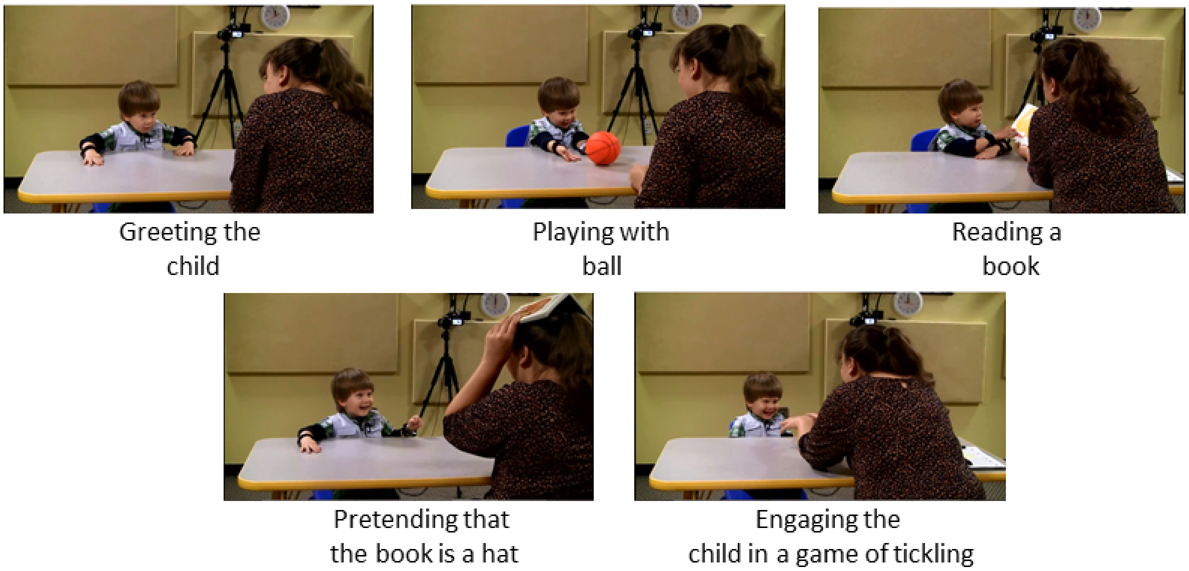
3.2 Multi-modal Dyadic Behavior Dataset

The Multi-modal Dyadic Behavior (MMDB) dataset [45] 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 who is assessed for interaction 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 3, 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.

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 whining 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



Greeting the child

Playing with ball

Reading a book

Pretending that the book is a hat

Engaging the child in a game of tickling

Figure 3: Stages of the dyadic interaction between child and examiner in the MMDB.

sessions is shown in Figure 4.

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 whining and they are present in a significant number compared to earlier studies with most of the laughter samples emanating during the tickling stage of the Rapid-ABC. The child's vocalizations are recorded using 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.

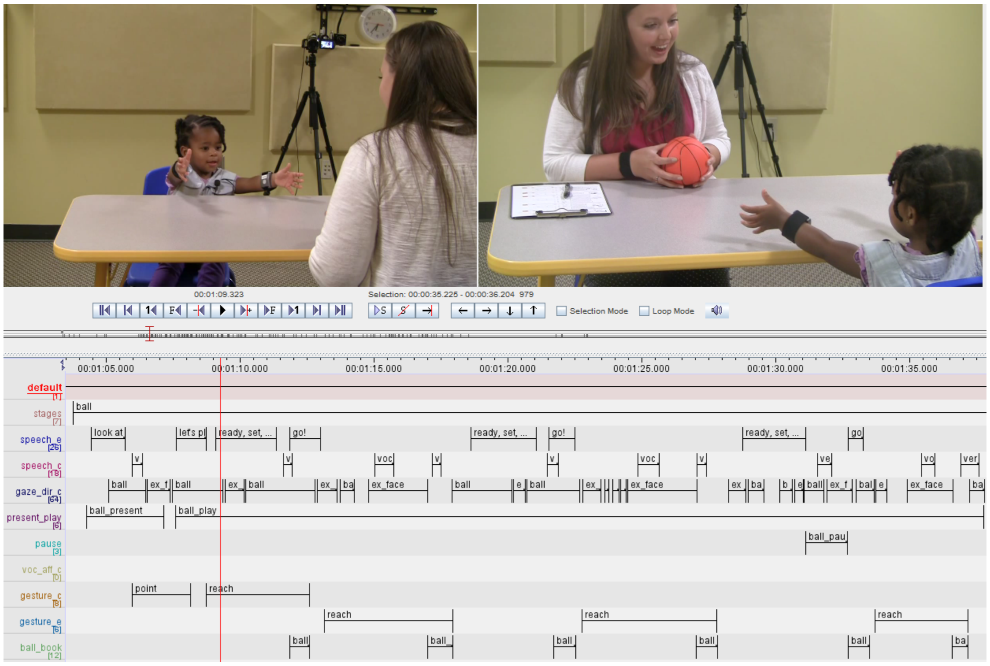


Figure 4: MMDB session annotations in ELAN.

3.3 FAU-Aibo Emotion Corpus

The FAU-Aibo Emotion Corpus (FAU-AEC) [37, 38], recorded at the Friedrich-Alexander University, Erlangen-Nuremberg, Germany, consists of recordings of adolescents during an interaction with Sony's pet AIBO robot. The corpus was recorded with 51 subjects (21 males and 30 females) whose ages ranged from 10-13 years. The robot was controlled by a human operator to perform a set of actions that would elicit naturalistic reactions from the subjects.

The significance of this dataset is that it has data that is annotated at the word and chunk level for children's speech and laughter which are of a spontaneous nature and has a significant number of samples ($N = 236$) of laughter.

3.4 Infant Brain Imaging Study

A set 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 data has been provided by research collaborators from the

University of North Carolina, Chapel-Hill (UNC, Chapel-Hill) and these are recorded at four different locations across the country. These sites including UNC, Chapel-Hill are Children's Hospital of Philadelphia, Philadelphia,PA, University of Washington, Seattle, WA, and Washington University in St. Louis, St. Louis, MO. There are 85 subjects in this study and the data is recorded at two time instances during the growth of the infant at 9 and 15 months of age. Data is collected from infants who are at low and high risk of ASD. The distribution of the subjects based on their risk factors is shown in Table 2.

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

The recordings of the child's interactions with their caregivers is 16 hours in length and were recorded using the Language Environment Analysis (LENA) device 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 5, has the ability to record single channel audio data at a sampling rate of 16 kHz.

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 child's and adult's vocalizations, cross-talk, background noise, electronic noise, and turn-taking events [46].

The LENA software does not provide a fine-grained analysis of the child's nonverbal vocalizations and does not provide timestamps of when the child laughed, cried, or produced any other kind of paralinguistic vocalizations. These important

Figure 5: LENA audio recording device used for infant vocal development analysis.

measures are key in understanding the social behaviors of children when they interact with their caregivers and given the fact that these are recordings of children who are high and low risk of ASD, the atypical characteristics of these events might be useful for the very early detection of ASD. For the data collected in the study, a research assistant at the Georgia Institute of Technology labeled the segments using various categories as enlisted in Table 3. 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, whining, 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 these are recordings which are recorded “in-the-wild” and constitute an important part in the scheme of validating models trained in laboratory environments, which are sound-treated and the vocalizations are produced in a completely different context, by testing them on the IBIS dataset. An important aspect of this dataset is also the presence of infant-directed speech and whether a causal relationship exists between adults’ speech directed towards infants and the paralinguistic event produced by the child.

3.5 Weill Cornell Medical College Database

The Weill Cornell Medical College (WCMC) corpus is a preliminary study of individuals with ASD to develop behavioral and neurophysiological measures sensitive to change in response to treatments. There are 16 families who have consented to taking part in the study and their children will participate in one week of home data collection and another at the Center for Autism and Developing Brain (CADB) in White Plains, NY. The study is meant to recruit children between the ages of 5 to 18 and may have limited (two or three phrases) vocabulary. The LENA device is used to record the audio data in both the locations. The data was annotated by two research assistants at WCMC using the annotation tool as described in the preceding subsection with the same set of labels.

3.6 Strange Situation

The Strange Situation protocol [47] is used for analyzing attachment behaviors of children with their caregivers. Attachment behaviors are observed in almost every child but an insecure attachment may result in developmental problems for the child. 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 separation events 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. This is indicative of their attachment behaviors with the caregiver and can be classified into one of three categories: secure, insecure avoidant, or insecure ambivalent. These attachment styles along with the classification criteria using crying [48] during the reunion episodes are shown in Table 4. The detection of crying is an important behavior considered in the scoring of this assessment.

Table 4: Classification criteria using crying in the Strange Situation protocol 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 that has been analyzed in this thesis was provided by research collaborators from 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, different age groups, and the type of crying produced in the Strange Situation consists of wailing while that of the MMDB is more of whimpering in nature.

3.7 Oxford Vocalizations Sounds Database

The Oxford Vocalizations (OxVoc) Sounds database is a collection of sound events of adults, infants, and domestic animals. These sound events are of a spontaneous nature and consists of events comprising of happy, sad, and neutral emotional states for humans. The adults' laughter and neutral speech events were obtained from video diary blogs and product reviews (primarily sourced from YouTube.com). This dataset will be used for validating our methods using novel acoustic features that captures the periodic structure of laughter for adults' speech.

3.8 SSPNet Vocalizations Corpus

The SSPNet Vocalizations Corpus (SVC) is a large collection of telephonic conversations (using a Nokia N900) of 120 adults (63 females and 57 males). The duration of the corpus is 8 hours and 25 minutes and the protocol consisted of participants having to talk about the Winter Survival Task. The data was annotated for laughter (N=2988) and fillers(N=1158). Again, the significance of this dataset is that it allows us to validate the predictive power of the novel acoustic feature which will be discussed in Chapter VI.

3.9 MAHNOB Laughter Database

The MAHNOB Laughter database [35] consists of recordings of 22 adults when they are shown short funny clips. This corpus is a multi-lingual corpus consisting of 12 males an 10 females. The average age along of males and females is 27 (standard deviation: 3) and 28 (standard deviation: 4) respectively. There are different types of laughter that are produced by the participants and this includes spontaneous and posed laughter. The recording protocol consists of showing several funny clips, used in previous research and from the internet, which lasted from a few seconds to two minutes. The subjects were also told to speak about a subject or interact with a friend or operator in English as well are their native language. The video recordings were done using a digital video recorder at 25 frames per second (fps) and it also has an in-built stereo microphone. A lapel microphone (single channel, sampling rate of 44.1 kHz) was used to record the audio in close proximity to the speakers. Thermal imaging was also used to record the data. The data was synchronized using a cross-correlation measure between the audio signals of the lapel and video recorder microphones. The data was annotated by a single rater and ELAN was used for annotation purposes.

The dataset has been used in this thesis to test the predictive power of the long-term syllable-level intensity features to detect laughter and for using OMRON's Okao library for detecting smiles. The fusion of multi-modal features from these two modalities can be used to improve the detection of laughter.

CHAPTER IV

DETECTION OF LAUGHTER IN TODDLERS' AND ADOLESCENTS' SPEECH

4.0.1 Laughter Detection in Children's Speech Using Spectral and Prosodic Features

There has already been work done on how to detect vocalizations such as laughter, in adults' speech. Detection of laughter in children's speech is less well explored and has important potential application in the clinical psychology domain. As described in Sections 2.4 and 2.5, previous research in analyzing paralinguistic events focused on a small set of acoustic features and with limited number of subjects. This section deals with the detection of laughter in the FAU-AEC using spectral and prosodic acoustic features from speech and laughter samples from children's vocalizations and verbalizations. The approach employed uses formant-based features that have not been explored in [37] and that have been found to have different articulatory kinematics for laughter in children's speech [49]. The information-gain-based feature selection technique was used in conjunction with a robust experimental setup, described in Section 4.0.1.3, to extract features with good class separability power.

4.0.1.1 Corpora

The datasets employed in the analyses are the Aibo Emotion Corpus (AEC) recorded at Friedrich-Alexander University (FAU), Erlangen-Nuremberg, Germany and the Multimodal Dyadic Behavior Dataset (MMDB) recorded at the Child Study Lab (CSL) at the Georgia Institute of Technology, Atlanta, GA.

FAU-Aibo Emotion Corpus The FAU-AEC corpus [37, 38] consists of interactions between children and Sony's pet robot Aibo. The vocalizations and verbalizations are spontaneous in nature as the children were led to believe that the robot was responding to their instructions. The laughter samples were annotated as well as the different types of laughter. These include speech which is modulated with laughter, voiced laughter, unvoiced laughter, and voiced-unvoiced laughter. In this stage of the research, the various types of laughter were treated as a single class. Sentences uttered by children were annotated as speech. The number of speech samples was 13478 and the number of laughter samples was 236. The research by Batliner et al. [37, 38] used samples of laughter which also had speech in them. We focused on extracting just the laughter portions from the samples as our focus was on building training models that will generalize well on to the MMDB dataset which had toddlers' vocalizations including speech and laughter. Also, for the purpose of duration normalization, we removed the silent portions in the speech samples of the FAU-AEC using a voice activity detector using Praat [51] as that would have resulted in features that would not have resulted in generalization when trying to match the conditions of the MMDB dataset. There is a discrepancy in the number of samples used by FAU and the current study by 16 events for speech and one for laughter. This is due to certain data being missing in the disseminated set. This discrepancy constitutes only 0.12% of the original dataset (13731 speech and laughter samples) and analysis differences are statistically insignificant when comparing results with FAU's.

Multi-Modal Dyadic Behavior Dataset The second dataset that was used was the MMDB [45]. In the context of the proposed research, the child may produce vocalizations in response to the activities and prompts made by the adult. Laughter is one of the key vocalizations that has been annotated and whose detection would aid in the diarization of the child's acoustic events and also help in analyzing the child's affective communication along with the level of engagement of the child with the adult. Twenty MMDB sessions were used for testing detection of laughter. The ages of the participants ranged from 15-29 months with a mean age of 22.45 months and a standard deviation of 4.62 months. The number of laughter and speech samples used for detection was 34 (17 for each class), with average duration of the laughter samples being 1.7 s, and average duration of a speech sample being 1.17 s. The differences between the datasets are the age groups, the context of the activity, and the presence of cross-talk in some of the samples with the adult talking in the background.

4.0.1.2 Feature Extraction and Selection

The open-source audio feature extractor, openSMILE [50], was used to extract 988 spectral and prosodic features using a 30 ms Hamming window with 10 ms overlap. The 52 acoustic features extracted using openSMILE are listed in Table 5.

Table 5: Spectral and prosodic acoustic features extracted using openSMILE.

Feature	**Number of Low-level Descriptors**
Intensity	2
Loudness	2
Mel-frequency cepstral coefficients	24
Pitch	2
Probability of voicing	2
Pitch envelope	2
Line spectral frequencies	16
Zero-crossing rate	2

Table 6: Statistical measures evaluated for each acoustic feature.

Statistical Measure
Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1 - 3, and 3 inter-quartile ranges.

The features, listed in Table 5, were extracted for each sample and 19 statistical measures, described in Table 6, were calculated for each acoustic feature. Along with these features, formant-based features were extracted using a 30 ms Hamming window with 10 ms overlap. The features were extracted using the widely-used speech analysis tool PRAAT [51], which used the Burg algorithm [52]. The first three formant frequencies, their respective bandwidths, the ratio of formant frequencies and bandwidths, the Euclidean distance between the formant frequencies, the Euclidean distance between formant bandwidths, and the Euclidean distance between the ratio of the formant frequencies were extracted, as shown in Table 15. The 14 statistical measures, described in Table 16, were measured, resulting in 294 formant-based frequencies. The resultant dimensionality of the feature space turned out to be 1282.

Table 7: Formant-based features extracted using Praat for the FAU-AEC dataset

Feature	**Number of low-level descriptors**
Formant frequency	3
Formant bandwidth	3
Ratio of formant frequencies	3
Ratio of bandwidths of formants	3
Euclidean distance between formant frequencies	3
Euclidean distance between formant bandwidths	3
Euclidean distance between ratio of formant frequencies	3

Table 8: Statistical measures evaluated for each formant-based feature.

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

One of the proposed research objectives was to evaluate the predictive power of features that are able to discriminate between laughter and speech in children's speech. Therefore, a feature selection algorithm based on information gain was used. Information gain is often used in decision trees [53] and measures the number of bits of information obtained for class prediction by knowing the presence or absence of a sample point in the classes.

Let $\{w_i\}_{i=1}^{M}$ be the set of classes, and for any attribute, $\{X_j\}_{j=1}^{N}$, which has been discretized to N levels, the information gain of the attribute is given in (3).

$$IG(w_i, X_j) = H(w_i) - H(w_i|X_j), \tag{1}$$

where $H(w_i)$ is the entropy of the class w_i and $H(w_i|X_j)$ is the conditional entropy of the class w_i given the discretized attribute X_j. Using the definition of entropy, (3) can be rewritten in terms of probabilities, as shown in (2):

$$IG(w_i, X_j) = -\sum_{i=1}^{M} Pr(w_i) log_2 Pr(w_i)$$
$$+\sum_{j=1}^{N}\sum_{i=1}^{M} Pr(X_j)Pr(w_i|X_j)log_2 Pr(w_i|X_j) \tag{2}$$

The information gain for each feature is evaluated and ranked in increasing order. The reduction in the dimensionality of the feature set is described in the next section.

4.0.1.3 Experimental Design

The experimental setup in [37] used 250 random sub-samples from the 13494 available samples of speech. A speaker-independent validation approach was used to select the features. In that approach, the sample points from one speaker were held out and a correlation-based feature selection was performed using the sample points from the remaining 50 speakers. Finally, the intersection of the features selected for 51 speakers was obtained, which resulted in a reduced feature set of 30 acoustic features.

Considering the large number of samples annotated as speech and the relatively small number selected (250), the previous method does not take into account the various levels of intonation in speech produced by the subjects in the study and this might not be captured using a small subset of speech samples. Five sets of 250 random sub-samples of speech were used. The analysis pipeline is shown in Figure 6. After the features have been extracted from the five different sets, as described in Section 3, and concatenated with the features from the laughter samples, feature selection based on ranking according to information gain was performed for each of the five sets. The number of features to be ranked according to the information gain was set to 100 for each set, and then the intersection of the features was obtained for the five sets. This process resulted in a reduced feature set of 30 spectral and prosodic features which are listed in Table 9.

Table 9: Acoustic features selected using feature selection based on information gain and experimental setup using five sets of data.

Feature	**Number of features selected**
Probability of voicing	12
Pitch	5
Mel-frequency cepstral coefficient	5
Line spectral frequency	3
First formant f. requency	5

4.0.1.4 Feature Interpretation

The selected features are important in the understanding of production of laughter in children's speech. The relevant features can be classified into three groups, pitch and voicing-based, spectral-based, and linear predictive coding (LPC)-based features.

Pitch and Voicing-Based Features Based on the findings of [54], the probability of voicing is greater in speech than in laughter for adults. This fact has been supported for children's speech too [49]. This could be due to the vowel-consonant structure of

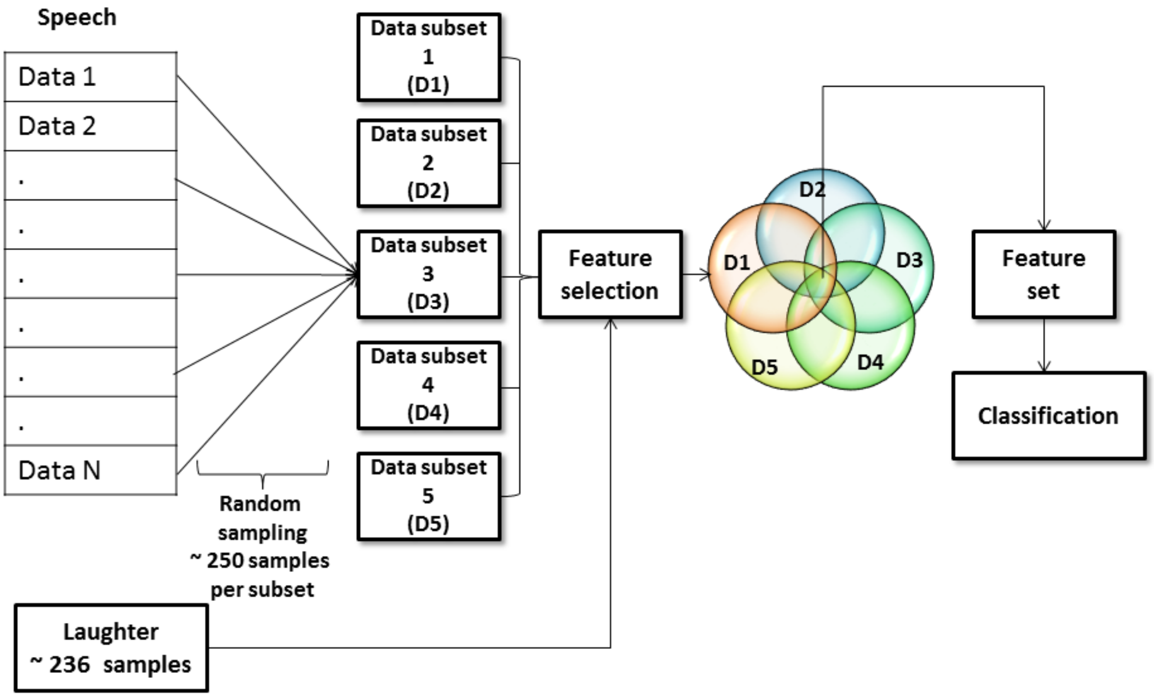


Figure 6: Diagrammatic representation of the methodology using five randomly sub-sampled sets of data along with the selection of features.

laughter. The work of [49] also suggests that the fundamental frequency (f_0) of children during laughter is high due to a high sub-glottal pressure and thin vocal folds [55]. The pitch and voicing-based features constitute nearly 60% of the features selected.

Spectral-Based Features The fourth MFCC was the only spectral-based feature that was selected using the experimental setup described in the previous section. The MFCC-based features, which emulate the psychoacoustical modeling of the human auditory system, have also been found to be prominent features in detection of laughter in adults' speech [56].

LPC-Based Features The LPC-based features consist of line spectral frequencies (LSF) and the first formant (F1) frequency. A pair of LSFs are the two resonant conditions that describe the vocal tract being either fully open or fully closed at the glottis [57]. In reality, the resonances occur when the glottis is neither fully open or fully closed and these are represented by formants as can be seen in Figure 7. Therefore, the LSFs and the formants share a symbiotic relationship. The findings of [49] suggest that laughter in children tends to have a high F1 owing to the fact of a more open mouth or a low jaw with young children exhibiting extreme kinematics with these articulators. These tend to become more controlled with development in age.

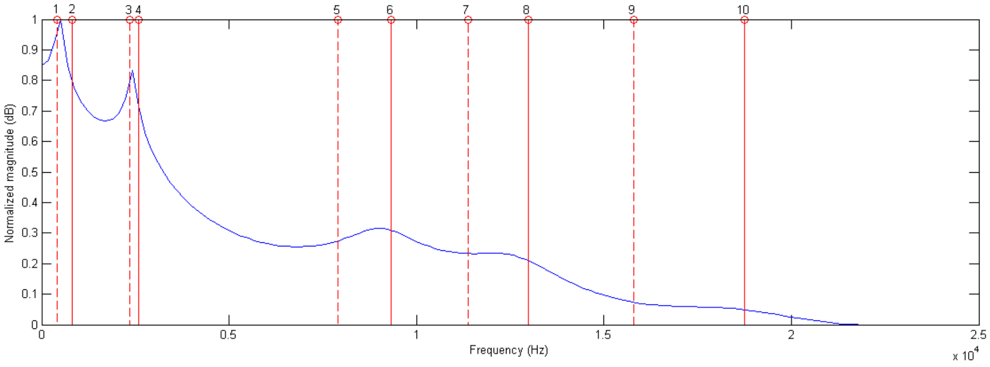


Figure 7: Spectrum of vocal tract response for the vowel /e/. The dashed and solid vertical lines represent the odd and even line spectral frequencies (LSF) respectively. The order of the LPC filter used is 10.

4.0.1.5 Results

For the purpose of classification, training models were developed on the reduced feature set using a variety of classifiers that include Gaussian mixture models using expectation-maximization (GMM-EM), multi-layer perceptrons (MLP), radial basis function neural networks (RBF-NN) and SVM with a multitude of kernels. The classification was performed using WEKA [58], an open-source machine learning software. The results using the various classifiers for a 10-fold cross-validation are shown in Table 10. The results indicate consistent accuracy for the five sets of data.

Table 10: Classification results using a 10-fold cross-validation scheme with various classifiers with average accuracy and standard deviation over the five sets of data.

Classifier	Accuracy (mean ± standard deviation)
MLP	95.04 ± 2.67%
RBF-NN	95.44 ± 2.70%
SVM (Linear kernel)	95.30 ± 2.68%
SVM (Polynomial kernel, degree=2)	95.82 ± 2.27%
SVM (RBF kernel)	95.96 ± 2.28%
GMM-EM	94.16 ± 3.25%

To evaluate the predictive nature of the selected features, the problem was treated as an unsupervised problem and clustering using GMM-EM and k-means was performed. The results are shown in Table 11. The error rate indicates that the features have robust predictive power.

Table 11: Clustering with GMM-EM and k-means with average error rate and standard deviation over the five sets of data.

Clustering Algorithm	Error rate (mean ± standard deviation
k-means	7.19 ± 3.67%
GMM-EM	5.71 ± 3.16%

To compare the proposed research work with the baseline results [37], the testing evaluation of a leave-one- speaker-out validation was performed. This validation was performed to ensure speaker independence. The classifier used for testing is an SVM with a quadratic kernel (degree = 1.65) and a complexity parameter (C=0.005). SVM was chosen for its superior generalization properties [59].

Table 12: Classification results of FAU using a support vector machine (SVM) on the FAU-AEC dataset.

	Predicted Speech	Predicted Laughter
True Speech	11054	2440
True Laughter	38	199

The accuracy of FAU's classification scheme is 81.95% and the average accuracy per class is 82.95% as shown in Table 12. The accuracy of the classification scheme is **94.43%** and the average accuracy per class is **94.46%** as shown in Table 13.

Table 13: Classification results using a support vector machine (SVM) with a polynomial kernel of degree = 1.65 and a complexity parameter (C = 0.005) on the FAU-AEC dataset with the proposed experimental design.

	Predicted Speech	**Predicted Laughter**
True Speech	12726	752
True Laughter	13	223

The results shown in Tables 12 and 13 indicate that the proposed method outperforms the baseline results **12.48%** in terms of absolute improvement. The results in Table 13 also indicate an equal error rate of **5.54%** for the classes of laughter and speech as this takes into account the huge imbalance between the classes.

An attempt was made to check if the models trained using the FAU corpus generalize to other datasets. This was done by testing on the MMDB dataset, described in Section 2 of the paper. Testing was performed on a relatively small number (17) of data points for each class. Again, an SVM was used with a linear kernel and the results are shown in Table 14.

Table 14: Classification results of the proposed research using a support vector machine (SVM) with a linear kernel and a complexity parameter (C = 1) trained on the data from the FAU-AEC dataset and testing on the MMDB dataset.

	Predicted Speech	**Predicted Laughter**
True Speech	12	5
True Laughter	5	12

The accuracy of the classification scheme both overall and per class was **70.58%**, which is significantly above chance (50%).

The results indicate a moderate generalization of the trained models on to other datasets. It is likely that for the lower than expected accuracy was due to differences in the age groups of the children in both the datasets. Our study has in many cases shown large acoustic and age differences, and there are instances of cross-talk in the MDBD dataset, which is not present in the FAU-AEC dataset and the speech samples in the MMDB dataset used for testing consisted of vocalizations, such as whining and crying along with other verbalizations whereas the FAU-AEC speech models were trained on intelligible speech. Nevertheless, the results on the MMDB dataset show our analyses generalize somewhat over highly mismatched conditions.

4.0.1.6 Conclusions

The research in this section was an attempt to analyze paralinguistic events in adolescents' speech using acoustic features. The experimental setup involving the use of randomly selected subsets of the FAU-AEC data captured the variations of the vocalizations of the database. The use of formant-based features was explored to discriminate between speech and laughter, and it was found found that the articulatory kinematics in the vocal tract during speech and laughter possess information to discriminate between them. In answer to the question of how generalizable the methods might be, models trained on a disjoint dataset, with subjects different in age, with different activity contexts, and with different amounts of cross-talk, all showed detection results significantly better than chance. We conclude that the proposed methods are using cues that are general to the task, and not specific to any one data set.

4.1 Detection of Laughter in Children with Autism Spectrum Disorder in Various Recording Environments

In this section of the research, the Weill Cornell Medical College (WCMC) corpus has been used for the purpose of detecting laughter in children with ASD. The uniqueness of this experiment is to test the generalization of features extracted from data recorded in a clinical setting and testing it in noisy environment home recordings.

4.1.1 Corpora

4.1.1.1 Weill Cornell Medical College Corpus

The dataset used in the detection of laughter in adolescents' and toddlers' speech with ASD was the WCMC which consists of home and clinic recordings of 16 children (aged 5-18 years of age) on the autism spectrum. The child's speech segments were labeled as speech, laughter, whining, crying, and other vocalizations for both the recording settings. For the clinic or baseline recordings, the number of laughter samples was 132 with a mean duration of 0.99 s and for the non-laughter segments (all other segments of child's speech other than laughter), the number of samples was 3293 with a mean duration of 1.09 s. In the home recordings, the number of laughter samples was 146 with a mean duration of 0.99 s and for the non-laughter segments the number of samples was 3537 with a mean duration of 1.13 s.

4.1.2 Feature Extraction and Selection

As described in Section 9, openSMILE was used to extract the baseline spectral and prosodic features. The features, listed in Table 5, were extracted for each sample and 19 statistical measures, described in Table 6, were calculated for each acoustic feature. Along with these features, formant-based features were extracted using a 30 ms Hamming window with 10 ms overlap. The features were extracted using the widely-used speech analysis tool PRAAT [51], using the Burg algorithm [52]. The first four formant frequencies and their respective bandwidths, along with the delta and delta-delta features were extracted as shown in Table 15. The 14 statistical measures, described in Table 16, were measured, resulting in 336 formant-based frequencies. The resultant dimensionality of the feature space turned out to be 1325.

As described in Section 4.0.1.2, the information gain criterion was used to select the features that are informative about detecting laughter. The information gain for each feature is evaluated and the top 100 features were selected.

Table 15: Formant-based features extracted using Praat for the WCMC dataset

Feature	Number of low-level descriptors
Formant frequency	12
Formant bandwidth	12

Table 16: Statistical measures evaluated for each formant-based feature.

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

4.1.3 Methodology

The experimental setup consists of building training models using the baseline recordings and testing it on the home recordings. In order to prevent overfitting to the majority class (non-laughter), we decided to randomly select 500 samples of it which would give sufficient diversity in terms of the type of vocalizations produced. The top 100 features selected using the information gain criterion is show in Table 17

Table 17: Acoustic features selected using feature selection based on information gain using the WCMC dataset.

Feature	Number of features selected
Probability of voicing	5
Pitch	23
Zero-crossing rate	5
Loudness and Intensity	10
Mel-frequency cepstral coefficient	32
Line spectral frequency	19
First and second formant frequencies and bandwidths	6

4.1.4 Results

For the purpose of classification, training models were developed on the reduced feature set and a cost sensitive classifier scheme was used. The cost sensitive matrix is given in (2) and is used to balance the instances in each class.

$$C = \begin{bmatrix} 0 & 3.77 \\ 1 & 0 \end{bmatrix} \tag{2}$$

Using a linear kernel SVM as the base classifier, the results using a 10-fold cross-validation is shown in Table 18.

Table 18: Classification results of the 10-fold cross-validation using the baseline recordings for training models.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	451	49
True Laughter	21	111

The average accuracy is **88.9%** and the average recall is **88.9%**. These results indicate that given data recorded in relatively clean recording conditions it is possible to discriminate between laughter and non-laughter events in children with ASD.

In order to test the predictive nature of the models, we tested our models on features extracted using the home recordings. The results are shown in Table 19 The average accuracy is **90.6%** and the average recall is **90.6%**. These results are pretty consistent with the test set results and indicate that it is possible to detect laughter in noisy conditions given training models in clean environments.

Table 19: Classification results of the 10-fold cross-validation using the baseline recordings for training models.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	3220	317
True Laughter	30	116

4.1.5 Conclusions

The research in this section attempted to detect laughter in children with ASD using spectral and prosodic features. The research in this section is one of the first few studies which has attempted to detect laughter in children on the autism spectrum. The selected features have been shown to be predictive enough to detect laughter in not only clean recording conditions but also in noisy environments as well.

CHAPTER V

PARALINGUISTIC EVENT DETECTION IN TODDLERS' INTERACTIONS WITH CAREGIVERS

5.1 Introduction

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 be important markers in the very early detection of autism spectrum disorder (ASD) [11, 12], and the diarization of such events in extended recordings can be a useful aid in the diagnosis of developmental disorders [13, 14]. It can also be used to analyze children's communicative behaviors in social interactions with their caregivers. The main focus of our work is to detect laughter and fussing/crying in toddlers' speech using acoustic features. Laughter is primarily used to express positive affect and has been found to usually follow a state of anticipatory arousal, especially tickling [10]. Fussing/Crying could indicate that the child is upset or disinterested in the task being initiated by the caregiver in a dyadic setting.

In this part of the research, the Multi-modal Dyadic Behavior (MMDB) dataset, the Strange Situation [47] corpus, was used for the purpose of developing detectors for laughter, fussing/crying, and child's speech consisting of verbalizations and vocalizations. The spectral and prosodic features were extracted using openSMILE [50], Praat [51], and VoiceSauce [60]. A brute force method of extracting features from toddlers' speech has been explored compared to earlier methods of using heuristics, described in Section 2.5 based on the type of paralinguistic cues to be analyzed. This enables the study of the gamut of acoustic features that have previously been less explored for this type of analyses. A combination of wrapper and filter-based feature selection approaches to reduce the dimensionality of the feature set was employed. The main aim of the analyses in this section is to investigate the generalization properties of the selected features to datasets that are disparate in not only the age range, but also the type of fussing/crying samples.

5.2 Corpora

The datasets that have been employed in this study are the Multi-modal Dyadic Behavior (MMDB) dataset, described in Section 4.1.1 and a set of 10 practice Strange Situations that had been conducted in multiple laboratories and were nationally distributed by researchers at the University of Minnesota, Minneapolis, MN.

5.2.1 Multi-modal Dyadic Behavior Dataset

There were 35 sessions randomly selected, which constitutes the training data, for detecting the child's paralinguistic events (laughter and fussing/crying) and the speech. The test set consists of 11 sessions. The ages of the participants ranged from 15 to 30 months with a mean of 21.65 and a standard deviation of 4.84. For analysis, the focus was on the child's verbal behavior for detecting instances of laughter, fussing/crying, and speech which were annotated by two research assistants in the CSL. The number of samples along with the mean and standard deviation of the duration of the samples of laughter, fussing/crying, and speech of the training and test sets are shown in Table 20. Owing to the large number of samples of children's speech and to prevent overfitting of the training data, the speech class was balanced by randomly selecting 58 samples.

5.2.2 Strange Situation Dataset

Recordings were made during the Strange Situation procedure [47]. The procedure consists of eight 3-minute episodes including two separations from the mother, each

Table 20: Number of training and testing examples of MMDB dataset 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 (s) (mean±standard deviation)**
Training Set	Speech (before balancing)	501	1±0.87
	Speech (after balancing)	58	1.14±0.66
	Laughter	54	1.31±1.28
	Fussing/Crying	62	2.65±4.21
Testing Set	Speech	122	1.23±0.92
	Laughter	35	1.12±0.90
	Fussing/Crying	30	1.68±0.83

followed by a reunion [48]. The episodes are arranged in a manner to create a series of stressful situations for the infant. 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. The detection of crying is an important behavior considered in the scoring of this assessment. In this dataset, only the fussing/crying events were annotated (N=62). The mean duration of the samples was 4.35 seconds and the standard deviation was 4.62.

The type of fussing/crying differs in both the corpora. The subjects in the MMDB dataset usually whimper to indicate discomfort with the activities, while in the Strange Situation recordings, the subjects cry when they are separated from the mothers.

5.3 Feature Extraction

The acoustic features were extracted using the open-source audio feature extraction tool, openSMILE [50]. There were 57 low-level descriptors (LLD), shown in Table 21 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 22, computed from the LLDs for each sample. The dimensionality of the feature set using openSMILE was 6669 and is relatively larger in comparison to the feature set used for the analyses in Section 4.0.1.2.

Table 21: 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 22: 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.

The formant-based features were extracted using Praat [51] and the cepstral peak prominence (CPP) was extracted using VoiceSauce [60]. The first four formant frequencies, resonances in the vocal tract [61], and their respective bandwidths were extracted. The delta and delta-delta for the formant-based frequencies were also extracted. The CPP is an approximate measure of breathiness in speech and is computed by measuring the difference between the peak of the cepstrum and a linear regression line fitted to the cepstrum [62]. It also gives a measure of the periodicity of the signal. These features are shown in Table 23. The total number of low-level descriptors for formant and CPP-based features was 25. The statistical measures, shown in Table 24, were computed for these features and the dimensionality of the formant-based and CPP features was 350.

Table 23: Formant-based and cepstral peak prominence features.

Feature	**Number of low-level descriptors**
Formant frequency	12
Formant bandwidth	12
Cepstral peak prominence	1

Table 24: Statistical measures evaluated for formant-based and cepstral peak prominence 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

5.4 Feature Selection

A two-pronged method of using both filter and wrapper-based approaches was used for feature selection. This incorporates the advantages of evaluating the intrinsic properties of the dataset using the filter-based method and the ability to generalize well by avoiding overfitting using the wrapper-based method. There is the added benefit of reduction in computation by selecting the k top features using the filter-based method, and then performing a wrapper-based feature selection on the reduced dimensionality feature set. The wrapper-based approach employs the correlation-based (CFS) and information gain ratio (IGR) feature selection techniques.

5.4.1 Correlation-based Feature Selection

The CFS [63] method selects features that are highly correlated with the class and uncorrelated with each other. For a subset of features S which contains k features and c classes, let r_{cf} be the mean feature-class correlation and r_{ff} be the mean feature-feature correlation, then the heuristic merit M_s is computed as shown in (3),

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{3}$$

The CFS method evaluates the correlation between a feature (k=1) and the class. The correlation between the feature and the class is computed using the Pearson correlation coefficient.

5.4.2 Information Gain Ratio Feature Selection

The information gain ratio (IGR) [64] is the information gain normalized by the intrinsic information of the feature. The information gain measures the number of bits of information obtained for class prediction by knowing the presence or absence of a sample point in the classes [65].

Let $\{w_i\}\,_{i=1}^{M}$ be the set of classes and for any attribute, $\{X_j\}\,_{j=1}^{N}$, which has been discretized to N levels, the information gain of the attribute is given in (4).

$$IG(w_i, X_j) = H(w_i) - H(w_i|X_j), \tag{4}$$

where $H(w_i)$ is the entropy of the class w_i and $H(w_i|X_j)$ is the conditional entropy of the class w_i given the discretized attribute X_j. One problem with the information gain criterion is that it favors features with a large number of values [64] and sometimes leads to overfitting.

The intrinsic information of the feature is computed by measuring the entropy of the class as shown in (5).

$$IV(X_j) = H(X_j), \tag{5}$$

where $H(X_j)$ is the entropy of the feature.

Features with high intrinsic value are considered to be less useful in discriminating between classes. The IGR, shown in (6), reduces the bias towards multi-valued features.

$$GR(w_i, X_j) = \frac{H(w_i) - H(w_i|X_j)}{H(X_j)}, \tag{6}$$

The filter method which gives the highest accuracy when using the openSMILE features, since they form the majority of the features in the set, using a 10-fold cross-validation using an SVM with sequential minimization optimization (SMO) for the binary classification tasks and a multi-class one-class-against-all SVM for the tertiary classification task, was selected as shown in Table 25. These were used as the intermediate feature set for the wrapper-based method. For the binary classification tasks, the threshold for ranking and selecting the openSMILE features was 100 and for the tertiary classification, the threshold was 200. The higher threshold for the tertiary scheme would enable the multi-class one-class-against-all classifier to discriminate between one class and the other classes which are treated as a singular class. For the formant and CPP-based features, the threshold was 50 for the three schemes.

Table 25: Results of 10-fold cross-validation using a support vector machine (SVM) with linear kernel for the filter-based feature selection methods for the openSMILE features along with results for formant and CPP-based features.

Classification Task	**Wrapper-based Feature Selection**	**Accuracy**	
		openSMILE features	**Formant and CPP-based features**
Speech vs. Laughter	CFS	**81.3**%	79.5%
	IGR	75.9%	75.9%
Speech vs. Fussing/Crying	CFS	**83.3**%	67.5%
	IGR	78.3%	68.3%
Speech vs. Laughter vs. Fussing/Crying	CFS	68.4%	56.6%
	IGR	**70.1**%	60.1%

For the binary selection tasks, the CFS method is used to select the top 100 openSMILE and 50 formant and CPP-based features. For the tertiary classification, IGR is employed to extract the top 200 openSMILE features and 50 formant and CPP-based features. It is of interest to note, from Table 25 that the results are better than chance for both the feature sets for all the classification tasks.

5.4.3 Sequential forward selection

The sequential forward selection (SFS) employs an SVM with SMO and a linear kernel for the binary classification tasks. The tertiary classification scheme employs a Multi-class classifier using a one class-against-all SVM with SMO and a linear kernel. This method selects the feature which generates the highest accuracy in the feature set and iteratively adds features to the set until there is no more improvement in the accuracy. The methodology employed in this study is shown in Fig 8.

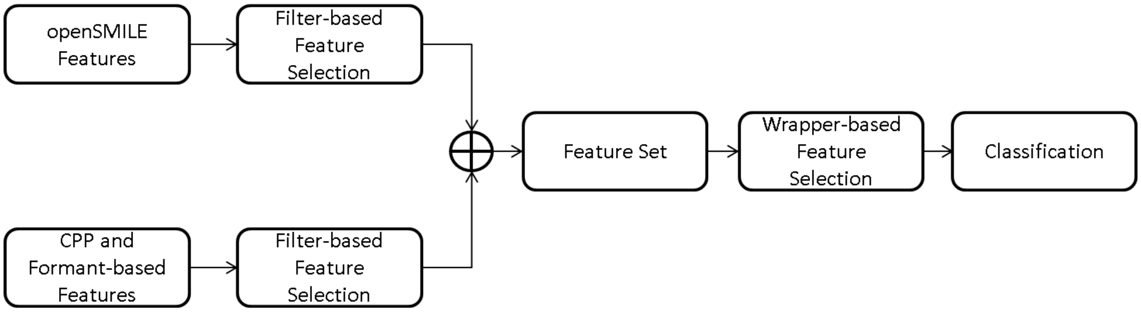

openSMILE Features
Filter-based Feature Selection
CPP and Formant-based Features
Filter-based Feature Selection
Feature Set
Wrapper-based Feature Selection
Classification

Figure 8: Method for selection of features using wrapper and filter-based feature selection methods for classification.

The purpose of the study is to understand which features are meaningful in discriminating between laughter, fussing/crying, and speech. As mentioned in Section 5.3, the two groups of features employed were the openSMILE and the formant and CPP-based features. These features were concatenated, after the filter-based feature selection, to form a 150-dimension feature set for the binary classification tasks and for the tertiary classification task, the dimensionality of the feature set was 250. The features were then processed to remove those with missing values and having less than 30% unique values. This was done to ensure that outliers did not affect the classification results. The SFS feature selection algorithm was used to further reduce the dimensionality of the feature set. The features and the number of statistical measures, selected using the wrapper and filter-based approaches, for the three classification schemes are shown in Table 26.

Table 26: Features selected for binary and tertiary classification tasks using combination of wrapper and filter-based features selection methods.

Feature	**Number of statistical measures**		
	Speech vs. Laughter	**Speech vs. Fussing/Crying**	**Speech vs. Laughter vs. Fussing/Crying**
Mel-frequency cepstral coefficient	1	6	3
Magnitude of mel-cepstrum	2	2	1
Pitch	-	2	1
Probability of voicing	1	-	1
Log Energy	-	-	1
Cepstral Peak Prominence	2	-	-
Spectral Rolloff	1	-	-
Spectral Centroid	1	-	-
Fourth formant bandwidth	1	-	-

5.5 Feature Interpretation

The MFCCs and the spectrum of the mel-spectrum constitute a major chunk of the features selected for the binary and tertiary classification tasks. The MFCCs, which aspects of human perception , have been found to discriminate well between adolescents' speech and laughter [65]. The pitch-related features, probability of voicing, pitch, and cepstral peak prominence, have also been found useful for all the classification tasks. These features are particularly useful for discriminating between speech and laughter, primarily due to the consonant-vowel structure of laughter. Whining or fussing has been found to exhibit higher pitch and varied pitch contours [66] when compared to adult-directed speech in children. The formant-based features, which were extracted using LPC analysis, weren't a part of the final feature set for detecting crying and the tertiary classification task. In comparison to the results in Section 4.0.1.6, the first formant frequency was not a part of the final feature set for detecting laughter. This can be attributed to the fact that children produce vocalizations which have a high pitch and the harmonics in the spectrum causes shifts in the positions of the formant frequencies.

5.6 Results

For the purpose of classification, training models were developed using the three reduced feature sets from the MMDB dataset. The classifier used is an SVM with SMO with a linear kernel and the open-source classification tool, WEKA [58]. A 10-fold cross-validation was performed on the three datasets and the results are shown in Table 27.

Table 27: Classification results using a 10-fold cross-validation scheme with support vector machine (SVM) with a linear kernel.

Classification Scheme	**Average Recall**	**Average Accuracy**
Speech vs. Laughter	**92.78%**	**92.85%**
Speech vs. Fussing/Crying	**88.49%**	**90.00%**
Speech vs. Fussing/Crying vs. Laughter	**76.43%**	**76.43%**

The results indicate that the laughter, fussing/crying, and speech can be discriminated robustly with the binary and tertiary classification schemes.

In order to test for the generalization of the results, a test set was devised consisting of 11 sessions selected randomly from the MMDB dataset. Seven sessions were used for the binary classification task of detecting laughter and the remaining 4 were used for detecting fussing/crying. The tertiary classification task used the combination of these. A grid search was performed by varying the complexity parameter, C.

Table 28: Classification results of test set of MMDB using SVM with linear kernel along with the complexity parameter, C, chosen using the grid search.

Classification Task	Accuracy	Precision	Recall	Complexity (C)
Speech (N=87) vs. Laughter (N=35)	**77.87%**	**74.44%**	**78.51%**	0.059
Speech (N=33) vs. Fussing/Crying (N=30)	**79.37%**	**80.72%**	**85.91%**	2.1
Speech (N=122) vs. Laughter (N=35) vs. Fussing/Crying (N=30)	**69.73%**	**66.15%**	**71.87%**	4.12

The results, shown in Table 28, indicate that the accuracy of classifying laughter and fussing/crying in children's speech is **77.87%** and **79.37%** respectively. For the tertiary scheme, the accuracy is **69.73%**. These results are significantly better than chance and show that the trained models generalize well to a test set from the MMDB dataset.

The Strange Situation dataset, as mentioned in Section 5.2.2, has only the fussing/crying events annotated. In order to test the trained models from the MMDB (Speech vs. Fussing/Crying) on this dataset, the features of the speech samples (N = 33) from the MMDB test set were concatenated with the features from the fussing/crying samples (N=62) of the Strange Situation dataset. This can be considered as a cross-corpus testing set. This gives a better sense of the generalization properties of the selected features and the trained models.

Table 29: Classification results of using trained models of MMDB (Speech vs. Fussing/Crying) and testing on a cross-corpus test set of MMDB and Strange Situation datasets using SVM with linear kernel and complexity parameter, C=2.1.

Classification Task	Accuracy	Precision	Recall
Speech(N = 33) vs. Fussing/Crying (N= 62)	**71.6%**	**73.4%**	**71.6%**

The results, shown in Table 29, indicate that trained models generalize well and are capable of discriminating between speech and fussing/crying with an accuracy of

71.6%. The findings are significant due to the different age groups of the participants, recording conditions, and the type of fussing/crying. The MMDB consists of fussing/crying or whimpering whereas the Strange Situation dataset consists of crying episodes. This indicates that the acoustic features are capable of not only capturing the characteristics of fussing/crying but also that of crying.

5.7 Conclusions

The research in this section has demonstrated the capability of robustly discriminating between children's speech, laughter, and fussing/crying. The combination of wrapper and filter-based features selection algorithms, which encapsulates the intrinsic properties of the dataset and generalizability, has the ability to select acoustic features that are relevant to laughter, fussing/crying, and children's speech. Through various experiments, it has been shown that these features have the predictive power to detect laughter and fussing/crying in children's speech. The selected features, trained on samples containing mainly fussing, are capable of robustly detecting crying when tested on to a database with a different age group.

CHAPTER VI

LONG-TERM FEATURES FOR DETECTION OF LAUGHTER IN CHILDREN'S AND ADULTS' SPEECH

6.1 Introduction

The previous two chapters focused on detecting laughter and crying in adolescents' and toddlers' speech using the baseline spectral and prosodic features. These features have been found to useful in detecting paralinguistic events to a significantly high degree of accuracy and generalized well to other datasets when trained on data recorded on subjects of a different age group and recording conditions. The features that were relevant to the tasks have also been found to be useful in detecting laughter in adults' speech and have also been used for speech recognition purposes. The techniques described above primarily use an agnostic process where features relevant to the database have been extracted and selected. The logical extension of this research is to investigate the use of features that can characterize the periodic nature of laughter using a long window.

Static short-term acoustic features have been widely employed to detect laughter in adults. These include prosodic features such as pitch, and energy, and spectral features such as mel-frequency cepstral coefficients [67]. These features are generally computed at the frame-level (30 ms Hamming window with 10 ms overlap) and capture the characteristics of stationary short-term windowed speech signal.

Laughter has been characterized as having a sonic structure which consists of a series of short vowel-like notes or syllables which are about 75 milliseconds long and repeated at regular intervals of about 210 milliseconds (4.76 Hz) apart [68]. Similar vowel sounds are used to define the structure of laughter and there are intrinsic constraints that define what constitutes laughter. For instance, the “ha-ha-ha” or “ho-ho-ho” structure would constitute laughter but not “ha-ho-ha-ho” which would sound unnatural. Research by Provine 1996 [68] has also shown that laughter in males has an average fundamental frequency of 276 Hz while that of females, about 502 Hz, is expectedly higher. The stereotypic structure of laughter is a result of the vocal apparatus and it is difficult to produce laughter which has a longer note duration than 75 milliseconds. An even longer inter-note interval makes laughter sound unnatural. Therefore, laughter can be considered to have a structural symmetry even though the symmetry does not exist in the amplitude which tends to decrease with the duration of the laugh. This can be attributed to the fact that humans run out of air ,and therefore a decrescendo in amplitude is observed.

Research [69] has shown that the use of long-term or syllabic level features conveys information about the rhythmic “ha-ha” structure of laughter. In that work, the Fast Fourier Transform (FFT) of the intensity contour is computed using a window size of 50 frames with a hop size of one frame and the magnitude in the bins between 4–6 Hz is summed. This *a priori* information about adults’ laughter in conjunction with other baseline acoustic features has been found to be useful in detecting laughter with an accuracy of 90% on the SSPNet Vocalisation Corpus.

This chapter would focus on building upon the work by [69] by developing a novel acoustic feature that captures the periodic properties in the intensity contour of laughter. Section 6.2 enlists the various databases consisting of children’s and adults’ laughter that have been used in this analysis. Section 6.3 describes the long-term syllable-level feature that has been developed and the succeeding section discusses the feature selection methods employed in this work.

6.2 Databases

The research in this chapter will focus on using long-term syllable-level features to detect laughter in children's and adults' speech. For this purpose, six datasets will be used. For the adults' laughter detection, the MAHNOB Laughter database, SVC, and OxVoc Sounds database will be used. For children's speech, MMDB, Strange Situation, and IBIS datasets will be analyzed. The MAHNOB Laughter dataset consists of vocalizations produced by adults while listening to funny clips and data has been annotated for speech and laughter. The number of samples and their duration is shown in Table 30.

Table 30: Samples of laughter and speech from the MAHNOB Laughter database with their respective mean and standard deviation of the duration.

Class	**Number of Samples**	**Duration (s) (mean±standard deviation)**
Speech	541	2.88 ± 2.18
Laughter	381	1.69 ± 2.45

For the SSPNet Vocalizations Corpus, which consists of telephonic conversations of adults, the number of samples along with the durations (mean ± standard deviation) is shown in Table 31.

Table 31: Samples of laughter and speech from the SSPNet Vocalizations Corpus with their respective mean and standard deviation of the duration.

Class	**Number of Samples**	**Duration (s) (mean±standard deviation)**
Speech(Filler)	1941	0.51 ± 0.25
Laughter	784	0.96 ± 0.72

For the OxVoc database, which consists of vocalizations produced by adults on social media, the number of samples along with the durations (mean ± standard deviation) is shown in Table 32.

Table 32: Samples of laughter and speech from the OxVoc Sounds database with their respective mean and standard deviation of the duration.

Class	**Number of Samples**	**Duration (s) (mean±standard deviation)**
Speech	30	0.91 ± 0.2
Laughter	30	1.5 ± 0

For detecting laughter in children's speech, we have used the MMDB, Strange Situation, and the IBIS dataset databases. The MMDB dataset, which consists of speech, laughter, and crying samples, has been used as the training data and the other three datasets are used as testing data. Table 33 shows the number of samples along with the durations (mean ± standard deviation) for all the datasets.

Table 33: 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 (s) (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

6.3 Long-term intensity-based feature

In this work, we have introduced a new measure to capture the long-term periodic structure of laughter using the energy or intensity contour. The work by [69] 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. This research will not use the *apriori* information about the frequency with which the sonic structure manifests 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 (7).

$$E[n] = \sum_{n=1}^{n} x[n]^2 \tag{7}$$

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

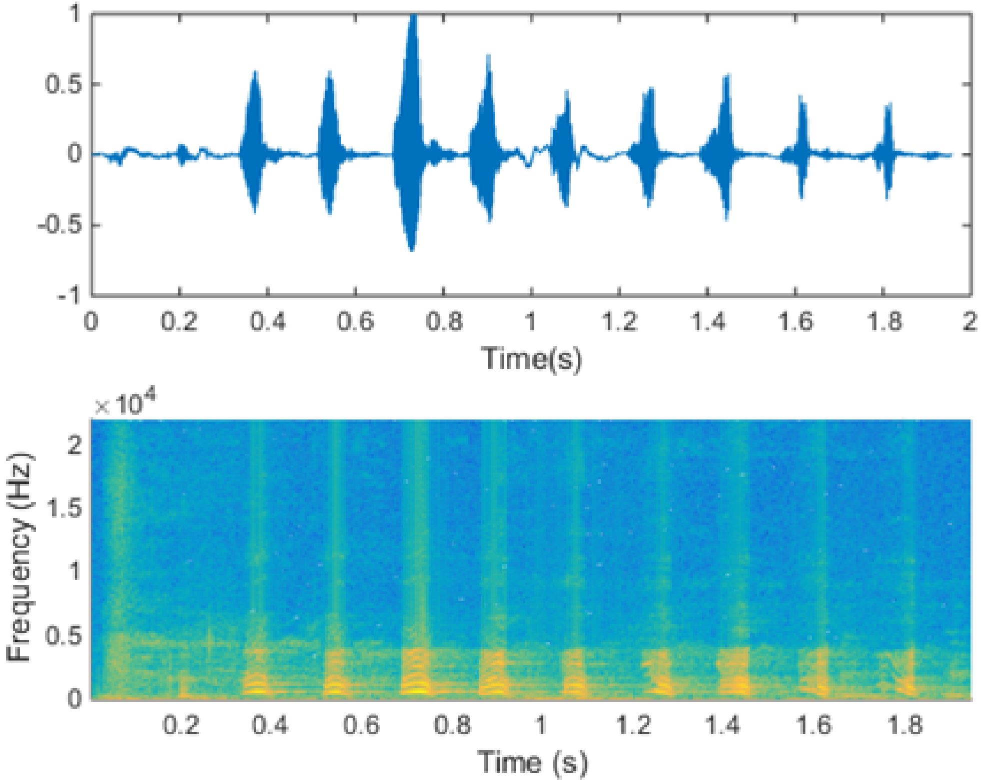


Figure 9: Waveform of laughter sample from the MAHNOB database along with the spectrogram displayed below it.

In Figure 9, the repetitive structure of laughter can clearly be seen in the spectrogram, while such a structure is not apparent for speech as seen in Figure 10. Using the intensity contour, the Hamming window length is again varied from 5 to 20 frames (in steps of 5) for adults' laughter and 5 to 45 frames (in steps of 4) for children's laughter

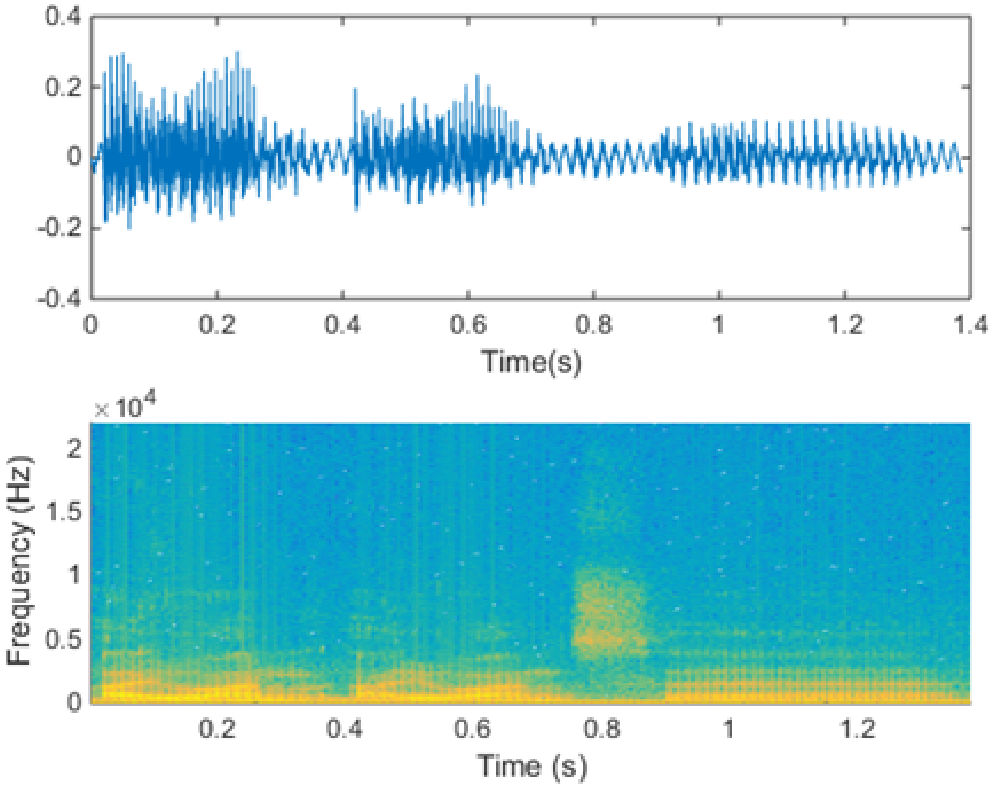


Figure 10: Waveform of speech sample from the MAHNOB database along with the spectrogram displayed below it.

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 6.4. From this syllable-level segment, the autocorrelation of the intensity contour is computed as shown in (8).

$$R_{xx}[j] = \sum_{n} x_n \bar{x}_{n-j} \tag{8}$$

Then, a polynomial regression curve is fitted to the one-sided autocorrelation function and the absolute error is computed between the curve and the autocorrelation function. The idea behind computing the error is that greater the periodic structure of the signal, which would be the case for laughter, higher would be the error than for speech. Since, the audio signals for adults' laughter detection we are using in this work consists of clean signals, we are fitting regression line to the autocorrelation function. On the other hand, since the children's audio signals might consist of noise or cross-talk, we are varying the degree, d, of the polynomial regression curve from 1 to 3. Also, for the children's speech we are using 4 different overlap window lengths ranging from 12.5% to 50% overlap whereas for adults' speech we don't user overlapping windows for computing the autocorrelation. This results in 42 low-level descriptors for adults' speech and 36 low-level descriptors for children's speech. There were 14 statistical measures computed from the features and these are shown in Table 34.

Table 34: 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

From this, we also obtained the delta and delta-delta features which resulted in an overall dimensionality of 168 features for adults' speech and 1512 features for children's speech.

6.4 Results

For the adults' laughter detection, we used the MAHNOB Laughter and SSPNet Voice databases for training our models. The OxVoc corpus consisting of adults' laughter and speech has fewer number of samples compared to the other two corpora and it would be pertinent to use the OxVoc samples for testing purposes.

Using the MAHNOB Laughter database, we trained our models using a linear kernel SVM and the results of the 10-fold cross-validation are shown in Table 35.

Table 35: Classification results of the 10-fold cross-validation using the syllable-level intensity features extracted from the MAHNOB Laughter database using a linear kernel SVM with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	507	34
True Laughter	23	358

The accuracy is **93.81%**, the average recall rate is **93.83%**, and the average precision rate is **93.49%**. Having trained the models on the MAHNOB database and testing on the OxVoc database, the confusion matrix is shown in Table 36.

Table 36: Classification results of testing on OxVoc dataset having trained on MAHNOB Laughter database using the syllable-level intensity features using a linear kernel SVM with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	30	0
True Laughter	1	29

The accuracy is **98.33%**, the average recall rate is **98.33%**, and the average precision rate is **98.38%**.

Using the SSPNet Vocalizations Corpus, we trained our models using a linear kernel SVM and the results of the 10-fold cross-validation are shown in Table 37.

Table 37: Classification results of the 10-fold cross-validation using the syllable-level intensity features extracted from SSPNet Vocalizations Corpus using a linear kernel SVM with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	1695	246
True Laughter	135	649

The accuracy is **86.01%**, the average recall rate is **85.05%**, and the average precision rate is **82.56%**. Having trained the models on the SSPNET Vocalizations Corpus and testing on the OxVoc database, the confusion matrix is shown in Table 38.

Table 38: Classification results of testing on OxVoc dataset having trained on SSPNet Vocalizations Corpus using the syllable-level intensity features using a linear kernel SVM with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	30	0
True Laughter	5	28

The accuracy is **91.67%**, the average recall rate is **91.67%**, and the average precision rate is **92.85%**.

The results show that the features are able to discriminate laughter from speech in adults' speech very robustly using various corpora with significantly good generalization across various corpora.

Having trained models on adults' speech, we wanted to check the viability of these features on children's speech and for this purpose we trained models using the MMDB dataset and tested the models on the Strange Situation and IBIS datasets. Using the same methodology for the adults' laughter detection, we ranked the top 200 features based on the correlation feature selection method. The results will be discussed in two categories, the first one will deal with classifying laughter against combinations of various categories (speech, whining, and non-laughter which consists of speech and whining) using only the top 50 features ranked by CFS syllable-level intensity features and the other 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 11.

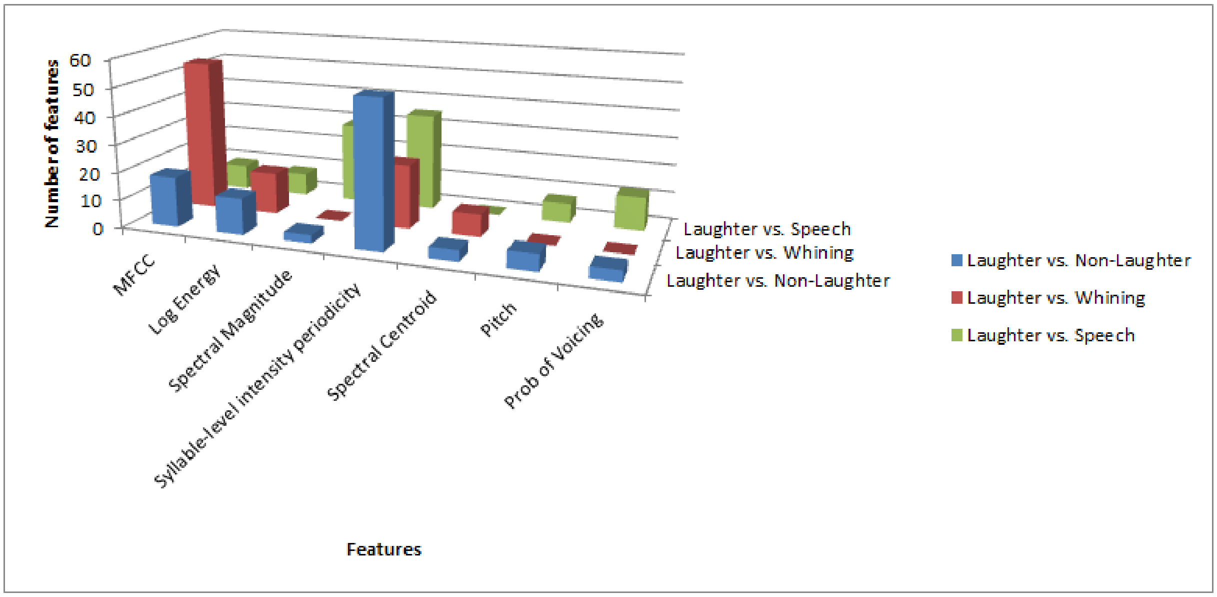


Figure 11: Features selected for the three classification tasks viz. speech vs. laughter, whining vs. laughter, and non-laughter vs. laughter

Using the MMDB corpora for training, the results of the 10-fold cross validation are shown in Table 40 for speech vs. laughter using the top 200 syllable-level features using CFS.

Table 39: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features using a linear kernel SVM ($C = 1$) with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	153	47
True Laughter	41	87

The accuracy is **73.17%** and the average recall rate is **72.23%**.

When the baseline features are used in conjunction with the syllable-level features and ranked using CFS, the results of the 10-fold cross-validation for speech vs. laughter are shown in Table 40.

Table 40: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features using a linear kernel SVM ($C = 1$) with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	169	31
True Laughter	19	109

The accuracy is **84.75%** and the average recall rate is **84.82%**.

Having trained models on the MMDB dataset, the model is tested on the test sets from the IBIS dataset, whose results are shown in Table 41, and on the Strange Situation corpus, whose results are shown in Table 42.

Table 41: Classification results of using the IBIS dataset as the test set with models training with the MMDB dataset using the baseline and syllable-level intensity features using a linear kernel SVM ($C = 0.0011$)with cost-sensitive learning for speech vs. laughter.

	Predicted Speech	**Predicted Laughter**
True Speech	434	76
True Laughter	7	41

The accuracy is **85.12%** and the average recall rate is **85.25%**.

Table 42: Classification results of using the Strange Situation corpus as the test set with models training with the MMDB dataset using the baseline and syllable-level intensity features using a linear kernel SVM ($C = 0.0012$) with cost-sensitive learning for speech vs. laughter..

	Predicted Speech	**Predicted Laughter**
True Speech	143	28
True Laughter	1	10

The accuracy is **84.06%** and the average recall rate is **87.26%**.

For laughter vs. whining, the results of the 10-fold cross-validation using the MMDB dataset with the syllable-level features are shown in Table 43.

Table 43: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features using a linear kernel SVM ($C = 1$) with cost-sensitive learning for laughter vs. whining.

	Predicted Whining	**Predicted Laughter**
True Whining	103	39
True Laughter	37	91

The accuracy is **71.85%** and the average recall rate is **71.81%**.

When the baseline features are used in conjunction with the syllable-level features and ranked using CFS, the results of the 10-fold cross-validation for whining vs. laughter are shown in Table 44.

Table 44: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features using a linear kernel SVM ($C = 1$) with cost-sensitive learning for speech vs. laughter.

	Predicted Whining	**Predicted Laughter**
True Whining	115	27
True Laughter	30	98

The accuracy is **79.25%** and the average recall rate is **78.77%**.

Having trained models on the MMDB dataset, the model is tested on the test sets from the IBIS dataset, whose results are shown in Table 45, and on the Strange Situation corpus, whose results are shown in Table 46.

Table 45: Classification results of using the IBIS dataset as the test set with models training with the MMDB dataset using the syllable-level intensity and baseline acoustic features using a linear kernel SVM ($C = 0.001$) with cost-sensitive learning for laughter vs. whining.

	Predicted Whining	**Predicted Laughter**
True Whining	341	80
True Laughter	9	39

The accuracy is **81.02%** and the average recall rate is **81.12%**.

Table 46: Classification results of using the Strange Situation corpus as the test set with models training with the MMDB dataset using the syllable-level intensity and baseline acoustic features using a linear kernel SVM ($C = 0.0005$) with cost-sensitive learning for laughter vs. whining.

	Predicted Whining	**Predicted Laughter**
True Whining	116	13
True Laughter	1	10

The accuracy is **90%** and the average recall rate is **90.41%**.

For laughter vs. non-laughter, the results of the 10-fold cross-validation using the MMDB dataset with the syllable-level intensity features are shown in Table 47.

Table 47: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features using a linear kernel SVM ($C = 1$) with cost-sensitive learning for laughter vs. non-laughter.

	Predicted Non-laughter	**Predicted Laughter**
True Non-laughter	262	80
True Laughter	35	93

The accuracy is **75.53%** and the average recall rate is **74.63%**.

When the baseline features are used in conjunction with the syllable-level features and ranked using CFS, the results of the 10-fold cross-validation for non-laughter vs. laughter are shown in Table 48.

Table 48: Classification results of 10-fold cross validation on the MMDB dataset using the syllable-level intensity features and baseline features using a linear kernel SVM ($C = 0.1$) with cost-sensitive learning for speech vs. laughter.

	Predicted Non-Laughter	**Predicted Laughter**
True Non	283	59
True Laughter	29	99

The accuracy is **81.27%** and the average recall rate is **80.04%**.

Having trained models on the MMDB dataset, the model is tested on the test sets from the IBIS dataset, whose results are shown in Table 49, and on the Strange Situation corpus, whose results are shown in Table 50.

Table 49: Classification results of using the IBIS dataset as the test set with models training with the MMDB dataset using the syllable-level intensity and baseline acoustic features using a linear kernel SVM ($C = 0.0002$) with cost-sensitive learning for laughter vs. non-laughter.

	Predicted Non-laughter	**Predicted Laughter**
True Non-laughter	771	160
True Laughter	11	37

The accuracy is **82.53%** and the average recall rate is **79.94%**.

Table 50: Classification results of using the Strange Situation corpus as the test set with models training with the MMDB dataset using the syllable-level intensity and baseline acoustic features using a linear kernel SVM ($C = 0.00025$) with cost-sensitive learning for laughter vs. non-laughter.

	Predicted Non-laughter	**Predicted Laughter**
True Non-laughter	250	50
True Laughter	1	10

The accuracy is **83.60%** and the average recall rate is **87.12%**.

The results indicate that the syllable-level features are capable of detecting laughter from speech, whining, 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.

6.5 Conclusion

The research in this section investigated the use of a novel acoustic feature that captures the periodic characteristics of laughter. The results of the adults' laughter detection show that the features are not only capable of detecting laughter to a high degree of accuracy and recall but can generalize well to other datasets. The features are also capable of discriminating laughter from other events in children's speech and when augmented with other acoustic features tends to generalize well to other datasets.

CHAPTER VII

MULTI-MODAL DETECTION OF LAUGHTER IN ADULTS' AND CHILDREN'S SPEECH

7.1 Combining acoustic and visual features to detect laughter in adults' speech

7.1.1 Introduction

Paralinguistic cues are non-phonemic aspects of human speech that are characterized by modulation of pitch, amplitude, and articulation rate [70]. These cues convey information about the affective state of the speaker and can be used to change the semantic content of a phrase being uttered. Research [71] has shown that 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 [5].

Charles Darwin, in his seminal work on emotions in animals, described laughter as a paralinguistic cue to primarily used to convey joy or happiness [72]. Laughter is a signal which consists of vowel-like bursts that has been found to be a highly variable signal. Research [73] has found 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" [73]. Furthermore, research has [74] 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. Research [75] has shown that in speed dating situations, women were rated to be flirting if they laughed while interacting with men.

Smiling is one of the most common facial expressions used while interacting with friends or peers [76]. 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 [77]. 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 [78].

There is limited understanding about the interaction between smile and laughter and one [79] 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 [22], this Chapter attempts to use the information about smiles to reduce false positives in detecting laughter using only the audio modality.

The chapter deals with the study of acoustic and visual features along with the fusion of the features to perform a multi-modal analysis of laughter in adults' and children's speech. The chapter briefly describes the data from the MMDB corpus used for training the models in Section 7.1.2. Section 7.1.3 describes the syllable-level acoustic and vision-based smile features extracted from the dataset. The methodology to create the feature set is described in Section 7.1.4 and the classification results are discussed in Section 7.1.5.

7.1.2 Database

For the adults' speech, the MAHNOB Laughter database, described in Section 3.9 was used. The database consists of spontaneous and posed laughter samples in a multi-lingual setting. The spontaneous laughter samples were elicited using a large collection of humorous video clips and the posed laughter samples were generated on command. The data was recorded using video and audio modalities. The video was recorded using digital video and thermal cameras. The audio was recorded using a lapel microphone as well as from the in-built microphone of the video camera. The corpus consists of 22 subjects but this study used data from the 15 subjects (9 males, 6 females) who provided consent for their recorded data to be published. For the purpose of our analyses, we used only the spontaneous laughter samples. We used audio from the lapel microphones and video recorded using digital video camera, all produced in a naturalistic manner. The samples used in the analyses along with the durations (mean ± standard deviation) for each class is shown in Table 51.

Table 51: Samples of laughter and speech from the MAHNOB Laughter database with their respective mean and standard deviation of the duration.

Class	**Number of Samples**	**Duration (s) (mean±standard deviation)**
Speech	541	2.88 ± 2.18
Laughter	381	1.69 ± 2.45

7.1.3 Acoustic and Vision-based feature extraction

Previous research has focused on using short-term prosodic and spectral acoustic features [80, 81, 82] for detecting laughter in corpora involving multi-speaker meetings. In computer vision, smiles have been described [83] as having a Facial Action Coding Units (FACS) coding consisting of lip corners pulled up and laterally along with varying levels of mouth opening. Also, visual features corresponding to head movements and facial expressions, were estimated using particle filtering tracking schemes [84].

This research focuses on using long-term acoustic features as described in Chapter VI and the use of a smile tracking software, described in the following subsections.

7.1.3.1 Audio Features

This research builds upon the idea of [69] by extracting features from the energy contour of the speech signal at the syllable-level but uses a different approach to measure the periodicity of the intensity contour by computing the autocorrelation of the intensity values.

As shown in Figure 12, the energy of the short-term speech signal is computed with a 30 ms Hamming window with a 10 ms overlap and the energy contour is extracted from it. The energy contour is smoothed using a median filter of window length of five samples to mitigate the effects of noise or extraneous vowels or consonants. The normalized one-sided autocorrelation function is computed for the smoothed energy contour using frames of varying length (5-20 frames with a step size of five). A linear regression line is fitted to the resulting autocorrelation function and the absolute error is computed from it. The error would be high for a periodic signal and vice-versa for a non-periodic signal.

7.1.3.2 Vision Features

The Omron Okao Vision library[1] is used to detect the subject's face, track the facial landmarks and extract visual features within the video signal. The Okao Vision library is a commercial facial analysis software that integrates face detection and tracking, facial landmark tracking, and face recognition. Face detection is performed at every frame, which is then processed by a face tracker to obtain the final position of a subject's face. The tracker provides a face identification (ID) number for each face in view and allocates a new ID when a new face is in view or the face tracker loses track due to occlusion. The *a priori* information of the first face, the subject's,

[1] http://www.omron.com/r_d/coretech/vision/okao.html

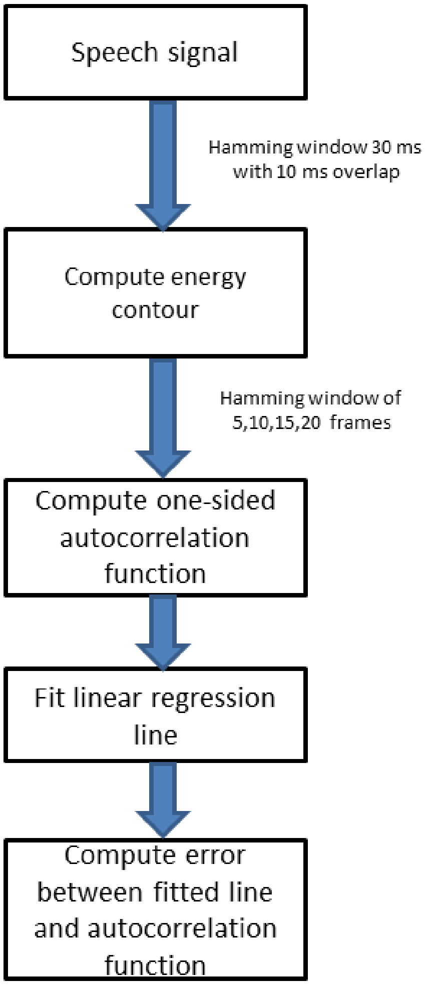

Speech signal
Hamming window 30 ms with 10 ms overlap
Compute energy contour
Hamming window of 5,10,15,20 frames
Compute one-sided autocorrelation function
Fit linear regression line
Compute error between fitted line and autocorrelation function

Figure 12: Schematic diagram of the feature extraction process of the energy contour at the syllabic level.

being tracked, is incorporated in the smile detection output generator to filter out instances of another person's face. With this pipeline, and due to a clear view of the subjects' faces in the dataset, we verify that the system is able to track all faces in the dataset.

Facial landmarks, e.g. mouth and eye corners, are then automatically estimated within a bounding box in each frame, based on image visual features. These landmarks are further used to estimate the smile degree of the subject, as well as its confidence. We use both the smile degree and its confidence as our vision features.

A median filter with a window of 10 frames was used to smooth the features. The frames when the face tracker did not detect the face due to occlusion all had their features set to zero. Two instances of when there is a smile and a lack of one are shown in Figures 14 and 13 respectively.

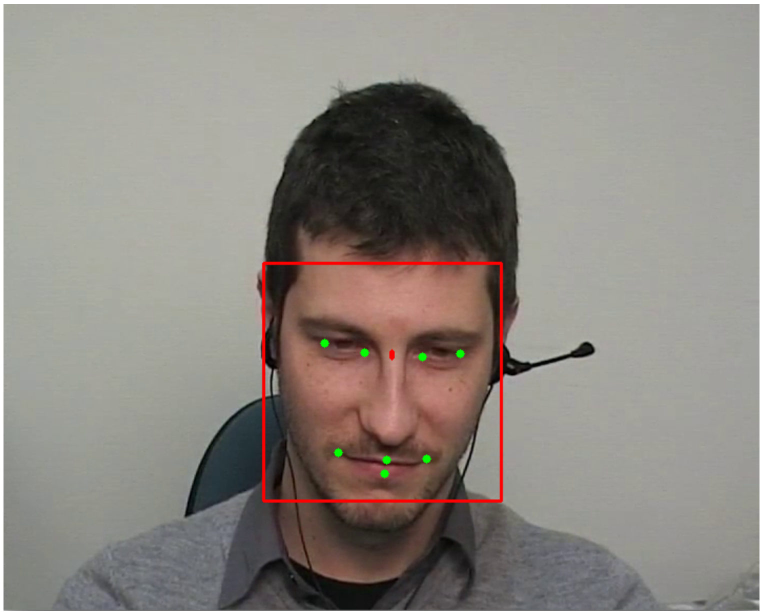


Figure 13: Analysis of smile detection when subject S001 does not smile.

7.1.4 Methodology

To account for the different frame rates of the audio and video modalities, we computed 14 statistical measures, as shown in Table 52 from speech and laughter segments.

Table 52: Statistical measures evaluated for acoustic and visual features.

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

For the audio features, the delta and delta-delta (first and second derivatives)

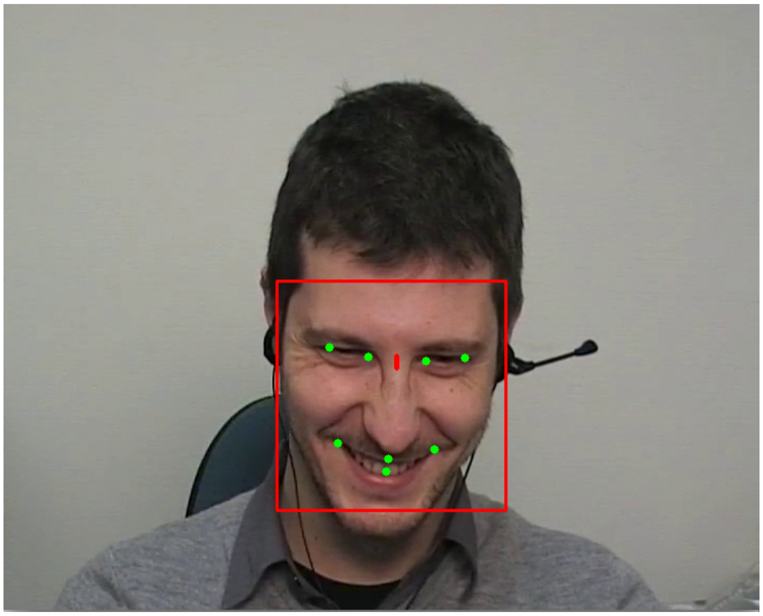


Figure 14: Analysis of smile detection when subject S001 smiles

measures were computed; for the visual features, only the delta measure was computed. Features that had not-a-number entries were removed and were normalized using z-score. For the audio features, the dimensionality was 168 and for the visual features, it was 48. The overall dimensionality of the audio-visual feature set was 216.

7.1.5 Results

A random forest (RF) classifier with 100 trees was trained using WEKA [58] on the speech and laughter samples. The cost-sensitive learning method [85] was used to account for the imbalance of the classes in the dataset. The cost matrix is given in (9) and balances the number of instances of each class.

$$C = \begin{bmatrix} 0 & 1 \\ 1.42 & 0 \end{bmatrix} \tag{9}$$

The evaluation process consists of performing a leave-one-speaker-out cross-validation with training models on $N-1$ speakers and testing on the N^{th} speaker. We performed three experiments on laugher classification: using only audio features, using only visual features, and the combination of both visual and audio features. Using only the audio features, the confusion matrix for speech vs. laughter is shown in Table 53. The accuracy is **93.06%**, the average recall rate is **93.11%**, and the average precision rate is **92.69%**. The results show that the syllabic features capture the periodic nature of laughter and are capable of discriminating between speech and laughter.

Table 53: Classification results of the leave-one-speaker-out cross-validation using only acoustic features with a random forest (RF) with 100 trees and cost-sensitive learning.

	Predicted Speech	**Predicted Laughter**
True Speech	502	39
True Laughter	25	356

The confusion matrix of just using the visual features is shown in Table 54 and the accuracy is **89.48%**, the average recall rate is **89.29%**, and the average precision rate is **89.09%**. The results are good considering that the features extracted using the vision modality can discriminate between speech and laughter which were annotated using only the audio modality. For fusion, the audio and visual features were concatenated and fed to the classifier. The confusion matrix of the early fusion method is shown in Table 55. The accuracy is **96.85%**, the average recall rate is **96.97%**, and the average precision rate is **96.6%**. The results are significant as the fusion of the features resulted in not only an improvement in the accuracy but a reduction in the false positive rates for laughter and speech.

Table 54: Classification results of the leave-one-speaker-out cross-validation using only visual features with an RF with 100 trees and cost-sensitive learning.

	Predicted Speech	**Predicted Laughter**
True Speech	489	52
True Laughter	45	336

7.1.6 Conclusion

The detection of laughter in adults' speech has been demonstrated with a very high level of accuracy. In this research, we explored the use of novel syllable-level acoustic

Table 55: Classification results of the leave-one-speaker-out cross-validation using fusion of audio and visual features with an RF with 100 trees and cost-sensitive learning.

	Predicted Speech	**Predicted Laughter**
True Speech	521	20
True Laughter	9	372

features to capture the periodic "ha-ha" structure of laughter. The use of visual features pertaining to smile were able to robustly discriminate between speech and laughter due to the fact that laughter and smiles are linked to each other. Our study demonstrates a significant improvement in accuracy and reduction of error rate by the proposed multi-modal early fusion of acoustic and visual features. These results are good enough to explore the use of the audio-visual features in naturalistic settings and specifically in situations where children interact with caregivers and peers, in order to analyze their social behaviors.

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

7.2.1 Introduction

The research in the preceding sections talks 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.

This section deals with the investigation of acoustic and visual features for the purpose of detecting laughter events in children's interaction with caregivers. Section

7.2.2 deals with the samples of the database along with the challenges encountered in extracting visual information from the videos. Section 7.2.3 describes the features and the feature selection techniques employed. Section 7.2.4 discusses the use of restricted Boltzmann machine (RBM) for the purpose of performing multi-modal fusion with a brief description of the architecture employed as well. The penultimate section has a discussion about the results and followed by what this study has learnt from performing multi-modal fusion of acoustic and visual features.

7.2.2 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 whining. 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).

7.2.3 Feature Extraction and Selection

The openSMILE features, described in Section 5.3, along with the syllable-level intensity features, described in Section 6.3, 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 as described in Section 7.1.3.2, and the features that were used for analyses were the smile strength. In this part of the research, there were two methods employed for feature selection. The first technique is the combination of the filter and wrapper-based techniques as used in Section 5.4 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 15, and thus each set of units is conditionally independent of the other.

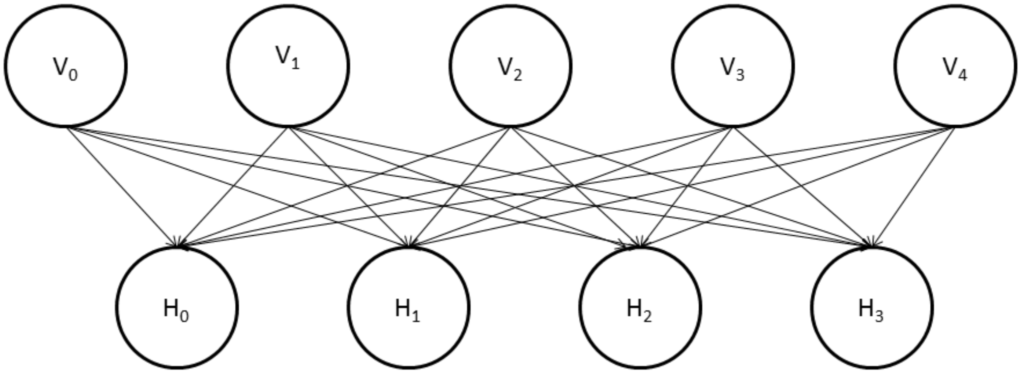


Figure 15: Structure of a restricted Boltzmann machine (RBM) with connections between visible layer, V, and hidden layer, H.

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 (10)

$$E(v,h) = -\sum\nolimits_{i,j} W_{ij} v_i h_j - \sum\nolimits_i a_i v_i - \sum\nolimits_j b_j h_j. \tag{10}$$

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 (11) and (12),

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum\nolimits_i W_{ij} v_i) \tag{11}$$

$$p(v_i = 1 \mid h) = \sigma(a_i + \sum\nolimits_j W_{ij} h_j) \tag{12}$$

where

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{13}$$

is the logistic function.

The probability that is assigned to every possible joint configuration (v, h) is given in (14),

$$p(v,h) = \frac{e^{-E(v,h)}}{Z} = \frac{e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}} \tag{14}$$

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

$$p(v) = \sum\nolimits_h p(v,h) \tag{15}$$

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

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_\infty \tag{16}$$

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 16.This sampling requires using the conditional distributions given in (11) and (12) which are easy to compute. The CD algorithm is given as,

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_k \tag{17}$$

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

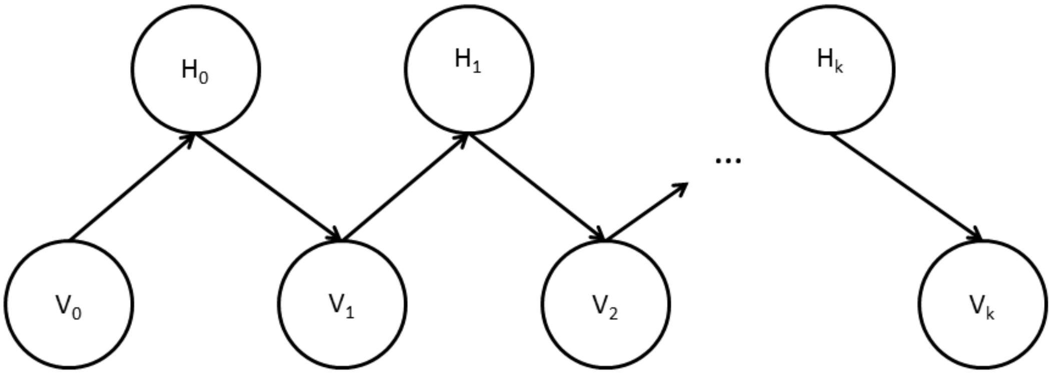


Figure 16: Working of the contrastive divergence (CD) algorithm between the hidden and visible units in an RBM.

visible units are treated as originating from a Gaussian distribution and the hidden units are binary. The equation of the energy function becomes,

$$E(v,h) = -\sum\nolimits_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum\nolimits_{i,j} \frac{v_i}{\sigma_i^2} h_j W_{ij} - \sum\nolimits_j b_j h_j. \tag{18}$$

The conditional probabilities of the visible and hidden units are modified as shown in (19) and (20).

$$p(v_i = v \mid h) = \mathcal{N}(v \mid a_i + \sum\nolimits_j W_{ij} h_j, \sigma_i^2) \tag{19}$$

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum\nolimits_i W_{ij} \frac{v_i}{\sigma_i^2}) \tag{20}$$

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

7.2.4 Methodology

Using the feature selection techniques described in the preceding subsections, we employed two methodologies for the multi-modal analysis. In the first part, as shown in Figure 17, 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 56 and include spectral centroid, syllable-level intensity, and smile confidence features.

Table 56: 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

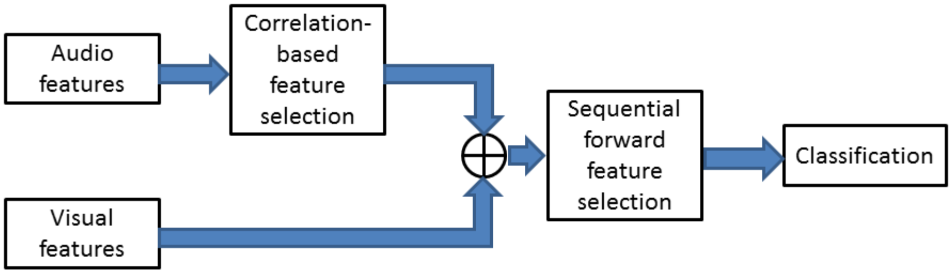

Audio features
Correlation-based feature selection
Visual features
⊕
Sequential forward feature selection
Classification

Figure 17: 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 [86]. Here, the lower layers learn the audio and video features separately followed by concatenating and feeding them to another RBM, as shown in Figure 18, 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 [87]. 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.

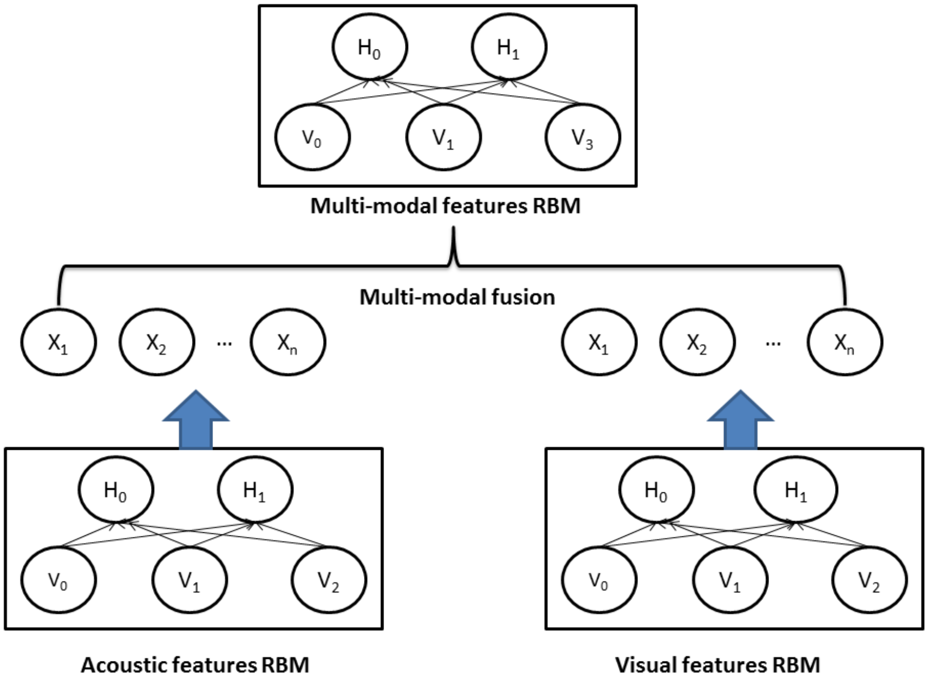


Figure 18: Architecture of the system employed for multi-modal laughter detection using RBMs.

7.2.5 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,

$$C = \begin{bmatrix} 0 & 1 \\ 1.81 & 0 \end{bmatrix} \tag{21}$$

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 57.

Table 57: Classification results of the 10-fold cross-validation using only acoustic features with a linear kernel SVM and cost-sensitive classification scheme.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	116	24
True Laughter	22	55

The result was an accuracy of **78.8%** and an average recall rate of **77.14%**. The recall rate of the laughter class is comparatively lower to the non-laughter class and the fusion with the visual features would help in reducing the number of missed detections.

Classification using only the video features, resulted in a confusion matrix for laughter vs. non-laughter as shown in Table 58.

Table 58: Classification results of the 10-fold cross-validation using only visual features with a linear kernel SVM and cost-sensitive classification scheme.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	111	29
True Laughter	12	65

The result was an accuracy of **81.1%** and an average recall rate of **77.14%**. The recall rate for the laughter class using visual features is significantly better than just using the acoustic features.

The combination of the features from both modalities followed by performing sequential forward selection (SFS) which yields in the reduced feature set as shown in Table 56 and resulted in the confusion matrix shown in Table 59.

Table 59: Classification results of the 10-fold cross-validation using visual and acoustic features as enlisted in with a linear kernel SVM (C=0.1) and cost-sensitive classification scheme.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	123	17
True Laughter	13	64

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.

For the multi-modal RBM architecture described in the previous section, using only the audio features with 40 hidden units in the RBM, resulted in a confusion matrix as given in Table 60.

Table 60: Classification results of the 10-fold cross-validation using RBM architecture with 40 hidden units for audio features with a classification scheme using linear kernel SVM with cost-sensitive learning.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	122	18
True Laughter	18	59

For the visual the features, the confusion matrix is shown in Table 61. This is the best result obtained using 10 hidden units in the RBM.

Table 61: Classification results of the 10-fold cross-validation using RBM architecture with 10 hidden units for visual features with a classification scheme using linear kernel SVM with cost-sensitive learning.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	113	27
True Laughter	16	61

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 confusion matrix is shown in Table 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

Table 62: Classification results of the 10-fold cross-validation using RBM architecture with 25 hidden units for visual and acoustic features with a classification scheme using linear kernel SVM with cost-sensitive learning.

	Predicted Non-Laughter	**Predicted Laughter**
True Non-Laughter	129	11
True Laughter	13	64

methodology.

7.2.6 Conclusions

This section 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.

CHAPTER VIII

CONCLUSIONS

Paralinguistic cues are known to convey the affective state of the speaker and especially in children are used as a form of communication with their caregivers.The detection of these events in hours of audio data could be potentially useful for clinicians for analyzing the atypical characteristics of these events in children with developmental disorders. The thesis has explored the use of acoustic features for the detection of these events across various datasets with different age groups, recording conditions, and protocols. Along with using features from only one modality, the thesis has also focused on using information from the vision modality to help improve the detection of laughter in children's speech.

The work in Chapters IV and V explored the use of spectral and prosodic acoustic features for the purpose of detecting laughter in children's speech. The research in these chapters focused on extracting features that characterize laughter and whining in adoloscents and toddlers speech. The research on detecting adoloscents laughter revealed that along with the baseline acoustic features, the formant-based features convey information about laughter using the feature selection algorithms employed. The significant finding of this study was the generalization of the results when models trained on adolescents' laughter were effective when tested on toddlers' laughter. Chapter IV investigated not only the predictive power of features useful for detecting laughter in adoloscents speech but also their generalization properties on toddlers speech as well. An important finding was the detection of laughter in children with ASD using both the baseline and formant-based features which encompasses a wide age range (5-18 years). The significant finding in this research was the generalization of the features selected using data recorded in relatively clean conditions in the laboratory to noisy data in home environments.

Chapter V is an extension from the work of detecting laughter in adoloscents speech to a completely different age group involving toddlers speech. The research in this chapter involves detecting instances of laughter and whining from toddlers speech while interacting with an examiner in a semi-structured interaction. The main finding of this work has been the use of dysphonation-related features (cepstral peak prominence) for classifying toddlers' laughter in their speech. These features likely capture the high pitch characteristics and breathy component of laughter. The important finding in this research has also been the generalization of the features on to a testing set consisting of data from subjects not part of the database as well as on to a different dataset involving infants in the Strange Situation protocol with a different recording environment.

An extension of this work is to investigate the development of features that would characterize laughters tonic structure. In Chapter VI, the thesis has focused on a novel syllable-level feature that uses time-series analysis to detect instances of laughter. These features were found to be robust enough to detect laughter in adults speech to a very high degree of accuracy and generalized well when applied to other datasets. In childrens speech, the features were moderately predictive to detect laughter and when augmented with baseline acoustic features, they were able to discriminate between childrens laughter from speech and whining and generalized well when applied to data recorded from infants in day long recordings with varying recording conditions and age group. The contribution of this research has been the ability to use long-term features which have been hypothesized in previous research to detect laughter.

The concluding portion of the thesis deals with the use of multi-modal information to detect laughter in childrens speech. The first part of Chapter VII used adults laughter database where the syllable-level intensity features were investigated along with computer vision-based smile related features. The important finding in this research was that using information from the vision modality was useful in detecting laughter in adults as well as childrens speech for events pertaining to the audio modality. The contribution of this work was in the investigation of the fusion of audio and video-related features and the subsequent improvement that was obtained in the detection of laughter.

REFERENCES

[1] D. K. Oller, R. E. Eilers, A. R. Neal, and H. K. Schwartz, "Precursors to speech in infancy: the prediction of speech and language disorders," *Journal of Communication Disorders*, vol. 32, no. 4, pp. 223–245, 1999.

[2] S. Nathani, D. J. Ertmer, and R. E. Stark, "Assessing vocal development in infants and toddlers," *Clinical linguistics & phonetics*, vol. 20, no. 5, pp. 351–369, 2006.

[3] D. Lerner, D. A. Adler, H. Chang, E. R. Berndt, J. T. Irish, L. Lapitsky, M. Y. Hood, J. Reed, and W. H. Rogers, "The clinical and occupational correlates of work productivity loss among employed patients with depression," *Journal of Occupational and Environmental Medicine*, vol. 46, no. 6, pp. S46–S55, 2004.

[4] K. L. Pike, "The intonation of american english." 1945.

[5] F. Poyatos, *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds.* John Benjamins Publishing, 1993, vol. 92.

[6] C. Darwin, *The Expression of the Emotions in Man and Animals.* Oxford University Press, 1998.

[7] J. Tepperman, D. R. Traum, and S. Narayanan, "" yeah right": sarcasm recognition for spoken dialogue systems." in *INTERSPEECH*, 2006.

[8] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[9] P. Ekman, R. J. Davidson, and W. V. Friesen, "The duchenne smile: Emotional expression and brain physiology: Ii." *Journal of personality and social psychology*, vol. 58, no. 2, p. 342, 1990.

[10] M. K. Rothbart, "Laughter in young children." *Psychological bulletin*, vol. 80, no. 3, p. 247, 1973.

[11] W. J. Hudenko, W. Stone, and J.-A. Bachorowski, "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*, vol. 39, no. 10, pp. 1392–1400, 2009.

[12] G. Esposito and P. Venuti, “Comparative analysis of crying in children with autism, developmental delays, and typical development,” *Focus on Autism and Other Developmental Disabilities*, vol. 24, no. 4, pp. 240–247, 2009.

[13] J. Hirschberg, “Dysphonia in infants,” *International Journal of Pediatric Otorhinolaryngology*, vol. 49, pp. S293–S296, 1999.

[14] J. Orozco and C. A. R. García, “Detecting pathologies from infant cry applying scaled conjugate gradient neural networks,” in *European Symposium on Artificial Neural Networks, Bruges (Belgium)*, 2003, pp. 349–354.

[15] F. J. Koopmans-van Beinum and J. M. van der Stelt, “Early stages in the development of speech movements,” *Precursors of early speech*, pp. 37–50, 1986.

[16] L. Roug, I. Landberg, and L.-J. Lundberg, “Phonetic development in early infancy: A study of four swedish children during the first eighteen months of life,” *Journal of child language*, vol. 16, no. 01, pp. 19–40, 1989.

[17] R. E. Stark, “Stages of speech development in the first year of life,” *Child phonology*, vol. 1, pp. 73–90, 1980.

[18] M. Zlatin, “Explorative mapping of the vocal tract and primitive syllabification in infancy: The first six months,” in *American Speech and Hearing Association Convention, Washington, DC*, 1975.

[19] D. K. Oller, E. H. Buder, H. L. Ramsdell, A. S. Warlaumont, L. Chorna, and R. Bakeman, “Functional flexibility of infant vocalization and the emergence of language,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6318–6323, 2013.

[20] S. M. Bell and M. D. S. Ainsworth, “Infant crying and maternal responsiveness,” *Child development*, pp. 1171–1190, 1972.

[21] L. Abou-Abbas, H. F. Alaie, and C. Tadj, “Automatic detection of the expiratory and inspiratory phases in newborn cry signals,” *Biomedical Signal Processing and Control*, vol. 19, pp. 35–43, 2015.

[22] M. Mehu and R. I. Dunbar, “Relationship between smiling and laughter in humans (homo sapiens): Testing the power asymmetry hypothesis,” *Folia Primatologica*, vol. 79, no. 5, pp. 269–280, 2008.

[23] J. A. van HOOFF, “A comparative approach to the phylogeny of laughter and smiling.” 1972.

[24] A. Scarpa, A. Raine, P. H. Venables, and S. A. Mednick, “Heart rate and skin conductance in behaviorally inhibited mauritian children.” *Journal of Abnormal Psychology*, vol. 106, no. 2, p. 182, 1997.

[25] W. Angst, "Basic data and concepts on the social organization of macaca fascicularis," *Primate Behavior: Developments in Field and Laboratory Research*, vol. 4, pp. 325–388, 1975.

[26] S. Preuschoft, *Laughter and smiling in Macaques: An evolutionary perspective.* Fac. Biologie, Rijksuniv. Utrecht, 1995.

[27] J. Van Hooff, "The facial displays of the catarrhine monkeys and apes." 1967.

[28] N. Blurton-Jones, "Non-verbal communication in children," *Non-verbal communication. New York: Cambridge University Press. p*, pp. 271–295, 1972.

[29] A. Fogel, G. C. Nelson-Goens, H.-C. Hsu, and A. F. Shapiro, "Do different infant smiles reflect different positive emotions?" *Social Development*, vol. 9, no. 4, pp. 497–520, 2000.

[30] D. S. Messinger, A. Fogel, and K. L. Dickson, "What's in a smile?" *Developmental Psychology*, vol. 35, no. 3, p. 701, 1999.

[31] C. K. Bainum, K. R. Lounsbury, and H. R. Pollio, "The development of laughing and smiling in nursery school children," *Child Development*, pp. 1946–1957, 1984.

[32] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–364.

[33] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[34] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "Avlaughtercycle," *Journal on Multimodal User Interfaces*, vol. 4, no. 1, pp. 47–58, 2010.

[35] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing*, vol. 31, no. 2, pp. 186–202, 2013.

[36] E. E. Nwokah, P. Davies, A. Islam, H.-C. Hsu, and A. Fogel, "Vocal affect in three-year-olds: A quantitative acoustic analysis of child laughter," *The Journal of the Acoustical Society of America*, vol. 94, no. 6, pp. 3076–3090, 1993.

[37] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On laughter and speech laugh, based on observations of child-robot interaction," *The Phonetics of Laughting, Trends in Linguistics. Mouton de Gruyter.*, 2010.

[38] A. Batliner, S. Steidl, and E. Nöth, "Associating children's non-verbal and verbal behaviour: Body movements, emotions, and laughter in a human-robot interaction," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 5828–5831.

[39] A. Zabidi, W. Mansor, L. Y. Khuan, R. Sahak, and F. Y. A. Rahman, "Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism," in *5th International Colloquium on Signal Processing & Its Applications, 2009. CSPA 2009.* IEEE, 2009, pp. 204–208.

[40] H. A. Patil, "Infant identification from their cry," in *Seventh International Conference on Advances in Pattern Recognition, 2009. ICAPR'09.* IEEE, 2009, pp. 107–110.

[41] P. Ruvolo and J. Movellan, "Automatic cry detection in early childhood education settings," in *IEEE International Conference on Development and Learning*, vol. 7, 2008, pp. 204–208.

[42] H. Zhang and G. Sun, "Feature selection using tabu search method," *Pattern recognition*, vol. 35, no. 3, pp. 701–711, 2002.

[43] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[44] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[45] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Scalaroff, I. Essa, O. Ousley, Y. Li, C. H. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, D. Lantsman, , J. Bidwell, and Z. Ye, "Decoding children's social behavior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.* IEEE, 2013.

[46] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[47] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, "Patterns of attachment. hills-dale," *NJ Eribaum*, 1978.

[48] E. Waters, "The reliability and stability of individual differences in infant-mother attachment," *Child Development*, pp. 483–494, 1978.

[49] C. Menezes and S. Diaz, "Phonetic and acoustic differences in child and adult laughter," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2517–2517, 2011.

[50] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia.* ACM, 2010, pp. 1459–1462.

[51] P. Boersma and D. Weenink, "Praat speech processing software," *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org.*

[52] J. P. Burg, "Maximum entropy spectral analysis." in *37th Annual International Meeting.* Society of Exploration Geophysics, 1967.

[53] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of international conference on new methods in language processing*, vol. 12. Manchester, UK, 1994, pp. 44–49.

[54] C. Bickley and S. Hunnicutt, "Acoustic analysis of laughter," in *Proc. Int. Conf. Spoken Language Process*, vol. 2, 1992, pp. 927–930.

[55] R. D. Kent and C. Tilkens, "Oromotor foundations of speech acquisition," *The International Guide to Speech Acquisition*, p. 2, 2007.

[56] L. S. Kennedy and D. P. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal.* National Institute of Standards and Technology, 2004, pp. 118–121.

[57] I. V. McLoughlin, "Review: Line spectral pairs," *Signal processing*, vol. 88, no. 3, pp. 448–467, 2008.

[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[59] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[60] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," *Energy*, vol. 1, no. H2, pp. H1–A1.

[61] J. C. Kim, H. Rao, and M. A. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction.* Springer, 2011, pp. 369–377.

[62] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, vol. 37, no. 4, p. 769, 1994.

[63] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[64] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, “Feature selection with high-dimensional imbalanced data,” in *IEEE International Conference on Data Mining Workshops, 2009. ICDMW’09.* IEEE, 2009, pp. 507–514.

[65] H. Rao, J. C. Kim, A. Rozga, and M. A. Clements, “Detection of laughter in children’s speech using spectral and prosodic acoustic features,” *Interspeech*, 2013.

[66] R. I. Sokol, K. L. Webster, N. S. Thompson, and D. A. Stevens, “Whining as mother-directed speech,” *Infant and Child Development*, vol. 14, no. 5, pp. 478–490, 2005.

[67] M. T. Knox and N. Mirghafori, “Automatic laughter detection using neural networks.” in *INTERSPEECH*, 2007, pp. 2973–2976.

[68] R. R. Provine, “Laughter,” *American scientist*, pp. 38–45, 1996.

[69] J. Oh, E. Cho, and M. Slaney, “Characteristic contours of syllabic-level units in laughter.” in *INTERSPEECH*, 2013, pp. 158–162.

[70] W. Apple, L. A. Streeter, and R. M. Krauss, “Effects of pitch and speech rate on personal attributions.” *Journal of Personality and Social Psychology*, vol. 37, no. 5, p. 715, 1979.

[71] J. Tepperman, D. Traum, and S. Narayanan, “” yeah right”: Sarcasm recognition for spoken dialogue systems,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[72] C. Darwin, *The expression of the emotions in man and animals.* Oxford University Press, 2002.

[73] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, “The acoustic features of human laughter,” *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[74] D. P. Szameitat, K. Alter, A. J. Szameitat, D. Wildgruber, A. Sterr, and C. J. Darwin, “Acoustic profiles of distinct emotional expressions in laughter,” *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 354–366, 2009.

[75] R. Ranganath, D. Jurafsky, and D. A. McFarland, “Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates,” *Computer Speech & Language*, vol. 27, no. 1, pp. 89–115, 2013.

[76] R. E. Kraut and R. E. Johnston, “Social and emotional messages of smiling: An ethological approach.” *Journal of personality and social psychology*, vol. 37, no. 9, p. 1539, 1979.

[77] U. Hess and P. Bourgeois, “You smile–i smile: Emotion expression in social interaction,” *Biological psychology*, vol. 84, no. 3, pp. 514–520, 2010.

[78] C. Meadows, *Psychological Experiences of Joy and Emotional Fulfillment*. Routledge, 2013.

[79] J. Lockard, C. Fahrenbruch, J. Smith, and C. Morgan, “Smiling and laughter: Different phyletic origins?” *Bulletin of the Psychonomic Society*, vol. 10, no. 3, pp. 183–186, 1977.

[80] S. Petridis and M. Pantic, “Audiovisual discrimination between laughter and speech,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5117–5120.

[81] L. S. Kennedy and D. P. Ellis, “Laughter detection in meetings,” in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*. National Institute of Standards and Technology, 2004, pp. 118–121.

[82] K. P. Truong and D. A. Van Leeuwen, “Automatic detection of laughter.” in *INTERSPEECH*, 2005, pp. 485–488.

[83] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

[84] I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for tracking facial features,” in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 97–102.

[85] C. Elkan, “The foundations of cost-sensitive learning,” in *International joint conference on artificial intelligence*, vol. 17, no. 1. Citeseer, 2001, pp. 973–978.

[86] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[87] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.