AUDIO DIARIZATION FOR LENA DATA AND ITS APPLICATION TO COMPUTING LANGUAGE BEHAVIOR STATISTICS FOR INDIVIDUALS WITH AUTISM

A Dissertation
Presented to
The Academic Faculty

By

Rahul Shivaji Pawar

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 2019

Copyright © Rahul Shivaji Pawar 2019

AUDIO DIARIZATION FOR LENA DATA AND ITS APPLICATION TO COMPUTING LANGUAGE BEHAVIOR STATISTICS FOR INDIVIDUALS WITH AUTISM

Approved by:

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

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

Dr. Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology

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

Dr. Rebecca M. Jones
Department of Psychiatry
Weill Cornell Medical College

Date Approved: August 20, 2019

ACKNOWLEDGEMENTS

This thesis has been possible due to contribution of many people during the course of my PhD. First and foremost, I would like to thank my advisor Mark Clements, for his continued support throughout my PhD. He showed immense patience and faith toward me, particularly in situations, when it might have been diffcult. I learned a tremendous amount due to him, and I hope to use some of his lessons in future.

I would like to thank Rebecca Jones, Meera Shobha, Meghan Swanson, Prof. Catherine Lord, and Prof. Joe Piven for being awesome collaborators. This work would not have been possible without these collaborations. My interactions with Rebecca and Meera were very inspiring, and most of my knowledge about autism research is due to them.

I am grateful to the members of my dissertation committee Prof. David Anderson, Prof. Elliot Moore, and Dr. Faramarz Fekri for their time, and suggestions, which made this thesis better.

I would also like to thank my friends and family, for their unconditional support through the lows and highs of graduate school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The objective of this dissertation is to develop diarization algorithms for LENA data and study its application to compute language behavior statistics for individuals with autism. LENA device is one of the most commonly used devices to collect audio data in autism and language development studies.

LENA child and adult detector algorithms were evaluated for two different datasets: i) older children dataset consisting of children already diagnosed with autism spectrum disorder and ii) infants dataset consisting of infants at risk for autism. I-vector based diarization algorithms were developed for the two datasets to tackle two scenarios: a) some amount of labeled data is present for every speaker present in the audio recording and b) no labeled data is present for the audio recording to be diarized. Further, i-vector based diarization methods were applied to compute objective measures of assessment. These objective measures of assessment were analyzed to show they can reveal some aspects of autism severity. Also, a method to extract a 5 minute high child vocalization audio window from a 16 hour day long recording was developed, which was then used to compute canonical babble statistics using human annotation.

CHAPTER 1
INTRODUCTION AND BACKGROUND

Autism Spectrum Disorder (ASD) is now one of the most prevalent developmental disorders among children in US. This is shown by the US Center for Disease Control and Prevention (CDC)'s Autism and Developmental Disabilities Monitoring Network (ADDM) report published in 2018, which concludes that ASD was prevalent among 1 in every 59 children aged 8 years in 2014 [1]. ASD refers to a group of neurodevelopmental disorders characterized by difficulties with social communication and interaction, and repetitive patterns of behavior. These include social impairment and communication issues such as difficulties with communication with people which involves mutual give-and-take, avoiding eye contact, limited verbal abilities, delayed speech and language abilities, difficulties understanding non-verbal cues such as gestures and body language. Some of the repetitive or unusual behaviors include flapping of their arms, rocking from side to side, or twirling.

Monitoring language and other acoustical behavior of infants at risk for autism, and of older children diagnosed with autism, could be useful, for early diagnosis in the first case, and for monitoring treatment changes in the second. In order to meet objectives of early diagnosis, and of tracking efficacy of treatments for individuals already diagnosed, it is crucial that automatic methods of monitoring language behavior are established for both controlled clinic conditions and for naturalistic home conditions . Accurate automatic methods would enable research studies with large number of participants, which in turn should enable more definitive conclusions about the hypothesis of any research study. If such automatic methods are accurate in natural home conditions, then, potentially, orders of more magnitude of audio data could be analyzed per participant, as there is a limit on the amount of data that can be collected in controlled settings. The limit exists because a) the participant will be willing to do only a certain number of visits to a clinic, and b) data collection during every clinic visit is resource intensive. Another advantage is that the role of caregivers in language development can be investigated because computation of language behavior measures based on interactions between caregivers and children would be possible. Thus, accurate automatic characterization of language and acoustical behavior of children in natural home conditions is of high significance to the autism research community.

The types of language and acoustic behavior measures that might be of interest in an autism research study depends upon its goals and hypothesis. For example, research studies involving infants might be interested in paralinguistic events such as crying, laughing, and screaming; canonical babble detection and characterization; quantity and quality of motherese etc. While, studies involving older children might be interested in the quantity and quality of child vocalizations, characterization of adult-child interactions using measures like conversational turns, response of children to adult questions etc. Automatic computation of these measures will help investigate their role in meeting the objectives of early diagnosis and monitoring treatment changes. To compute these measures, it is crucial that to be able to accurately determine when the child and adult are vocalizing in a given audio recording. The problem of who is speaking when is known as "Speaker Diarization" in the speech processing community. The current state of the art speaker diarization techniques are based on i-vector methods. These methods have performed well in traditionally studied domains in speaker recognition and diarization research such as telephony, broadcast news and meeting data. These domains usually have very large curated labeled databases available for training speaker recognition or diarization systems. However, the audio data for our studies is a lot more "in the wild" since it is collected in natural home environments using a single-channel recording device (LENA) attached to the clothing of children under study. Another challenge compared to the traditionally studied domains is that we have significantly less labeled data for training speaker recognition or diarization systems.

As mentioned earlier, the audio recordings that are to be investigated in the proposed research, were collected by a device called LENA, shown in Figure 1.1. It has recently become the primary method of collecting audio recordings for autism research studies. It is compact, light-weight, has a digital recording capacity of 16 hours, and affords 16 hours of battery life [2]. It is small enough to be worn even by infants, but is suitable for all ages given proper accommodations designed into clothing. One of its limitations is that it only makes a single-channel recording, which precludes the use of 2-channel methods for enhancing the signal. The supplier of the device also provides software to perform automated analysis of the audio recordings. The analysis software does segmentation based on speaker boundaries, classifies them into segments such as child vocalization, adult vocalization, silence etc., and uses this information to compute statistics. The set of analyses are very rich and go from low-level features, such as identifying when child and adult vocalizations occur, to high-level statistics, such as identifying the number of conversational turns and adult word count in an audio recording. However, these analyses are not accurate enough for meaningful analysis of data to be investigated in the proposed research.

Figure 1.3 shows the pipeline for automatic computation of language and acoustic measures from LENA data. The primary focus of this thesis is to develop methods to do diarization of LENA data for both older children and infants. Another focus of the research is to show that developed i-vector based diarization methods can be used to compute some of the language behavior statistics listed in stage 2 of the research pipeline 1.3. A detailed computation of all the statistics listed in stage 2 of the research pipeline and investigating their role in early detection of autism and monitoring treatment changes is beyond the scope of this work.

The thesis is organized as follows. *Chapter 2* describes the datasets analyzed in this work and GUI toolkit developed for annotation. It also evaluates the accuracy of LENA software for these datasets and shows that they are not not accurate for these datasets. *Chapter 3* describes the development of i-vector based diarization methods for two scenarios: 1) development of semi-supervised i-vector based diarization for the scenario when

Figure 1.1: LENA Device

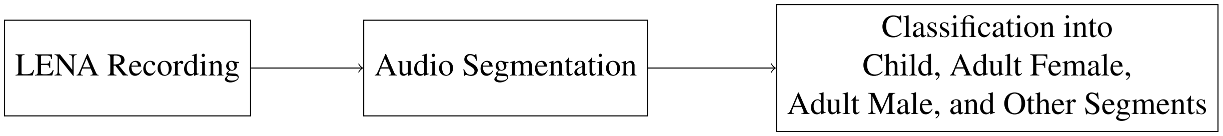
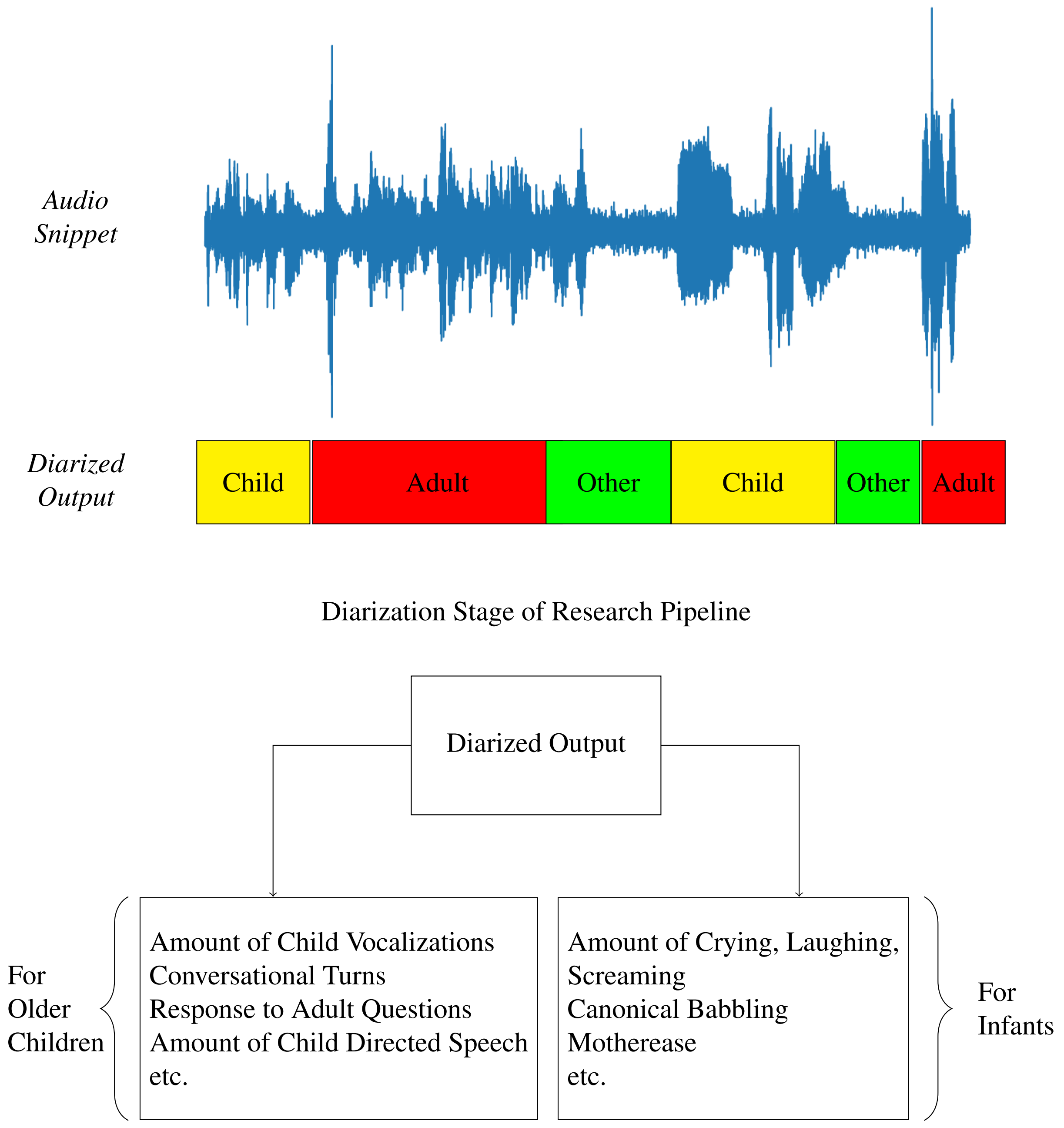


Figure 1.2: Diarization Steps for a LENA recording

some amount of labeled data is available for speakers present in the audio recording to be analyzed, and 2) development of unsupervised i-vector based diarization method for the scenario when no labeled data is available for speakers present in the audio recording to be analyzed. *Chapter 4* describes the application of semi-supervised child detectors to compute a metric called utterance rate, analyze its reliability across a period of 3 weeks and check if it can reveal some measure of autism severity. A method to obtain 5 minute high child vocalization audio window from 16 hour day long audio recording is described and computation of canonical babble ratio from this 5 minute audio window is shown. *Chapter 5* describes the contributions of thesis and illustrates possible future directions of research.



Diarization Stage of Research Pipeline

Language Behavior and Acoustic Characterization Stage of Research Pipeline

Figure 1.3: Pipeline for Automatic Computation of Language and Acoustic Measures

CHAPTER 2

DATASETS DESCRIPTION, ANNOTATION AND ANALYSIS OF LENA SOFTWARE

The child's acoustic environment in studies investigated in this work was recorded using a well engineered device called LENA. It is compact, light-weight, has a digital recording capacity of 16 hours, and affords 16 hours of battery life [2]. It is small enough to be worn even by infants, but is suitable for all ages given proper accommodations designed into clothing. One of its limitations is that it only makes a single-channel recording, which precludes the use of 2-channel methods for enhancing the signal. LENA device has become the primary instrument for collecting audio data in autism and other language development studies [3–15]. The list of papers cited is not exhaustive, but illustrates the wide usage of LENA device in autism and other language development studies. The supplier of the device also provides software to perform automated analysis of the audio recordings. The analysis software does segmentation based on speaker boundaries, classifies them into segments such as child vocalization, adult vocalization, silence etc., and uses this information to compute statistics. The set of analyses are very rich and go from low-level features, such as identifying when child and adult vocalizations occur, to high-level statistics, such as identifying the number of conversational turns and adult word count in an audio recording.

This chapter describes a) usage of LENA device and software in autism and language development studies, b) datasets investigated in this work, c) annotation method developed for datasets under investigation, and d) analyzes the accuracy of LENA software for datasets studied in this work.

2.1 Usage of LENA Device and Software in Autism and Language Development Studies

As mentioned earlier, LENA device has become one of the primary instruments for collection of audio data in autism research studies. [6] was one of the first studies to use automatic measures from LENA software for analysis. It concluded that it is feasible to compute stable automatic measures of vocal development from single day recordings. In [7], associations between the Adult World Count (AWC) measure, generated by LENA software and cognitive ability of preschoolers with ASD was analyzed. The study had 67 participants and data was collected during morning classroom routines. Results indicated that AWC was positively correlated with children's cognitive ability. In [8], quantitative and qualitative characterization of the school and home environments of 10 preschool children with ASD was carried out. Language samples were collected at 3-month intervals over the course of one year using LENA. For every sample, 15 minute segments were selected and transcribed using Systematic Analysis of Language Transcripts (SALT). These segments were then analyzed using LENA's AWC measure and SALT transcriptions. It concluded that there were significant differences between school environment and home environment for both AWC measure and for SALT transcription. However, no analysis was carried out to compare AWC measure with SALT transcriptions. In [10], stability and validity of two automatic methods of computing vocal development was tested. One method was the commercially available LENA software, while the other method was described as currently in development. The paper concluded that the vocal development measures computed by current LENA software are stable but do not correlate with future spoken vocabulary. However, their development software measures were stable and according to them, predicted future spoken vocabulary to a degree that was non-significantly different from the index derived from conventional communication samples. The study concluded that automated vocal analysis is a valid and reliable alternative to time intensive conventional manual methods of computing vocal measures. In [9], analysis of child-adult interaction during naturalistic day long LENA recordings of children with and without autism aged 8 to 48 months was carried out. It was shown that for both typically developing and children with autism, there exists a social feedback loop in which, adults are more likely to respond to child's speech like vocalizations than to child's non-speech like vocalizations. Also, it was shown that a child's vocalization was more likely to be speech related if child's previous speech like vocalization received an immediate adult response. The differences in social feedback mechanism between typically developing and children with autism were highlighted. It concluded that that such differences will influence language development over time.

LENA device is now also commonly used in non-autism language development research studies. In [4], examination of how audible and intelligible educator talk influenced infants under 2 year olds, who attended early childhood education and care (ECEC), was done. In [5], language experience was measured from home audio recordings of 36 children aged 4-6 years from diverse socioeconomic status. The study concluded that during a story-listening functional Magnetic Resonance Imaging (MRI) task, children who had experienced more conversational turns with adults exhibited greater left inferior frontal (Broca's area) activation, which significantly explained the relation between children's language exposure and verbal skill. In [11], relationships between the amount of language input and neural responses in English monolingual and Spanish-English bilingual infants was analyzed. In [12], influence of child-directed speech in two languages on language development of bilingual children was studied. In [13], investigation of whether children and caregivers modulate the prosody of their speech as a function of their interlocutor's speech was done. The study found small, but significant, effects of mothers and their children influencing each other's speech, particularly in pitch measures.

The above studies show that LENA is widely used as a means of collecting audio data for autism research studies and other language development studies. The above studies show that automatic computation of language and other acoustic behavior measures is actively researched as a tool for early autism detection and treatment monitoring for ASD.

However, the current automatic tool in the form of LENA software is not accurate enough for children of all age groups. In particular, it is shown in this chapter that the LENA software is not accurate for two datasets, namely *Older Children Dataset* and *Infants Dataset* studied in this work.

2.2 Datasets Investigated in this work

This section describes the two datasets for which i-vector based diarization methods were developed and applied to compute language behavior statistics for dataset 1. Dataset 1 termed as *Older Children Dataset* for the remainder of this dissertation had children with aged 5 to 13 years old. Dataset 2 termed as *Infants Dataset* had audio collected for infants at 9 month and 15 month age.

2.2.1 Older Children Dataset: Weill Cornell Study

Thirty-seven families were recruited through the Center for Autism and the Developing Brain (CADB) in White Plains, NY to participate in a study examining novel outcome measures. Participants (target-child) were 5-17 years old (31 boys), see Table 2 for participant demographics. The language level of the target-child varied from two to three word phrases to fluent speech. Weill Cornell Medicines IRB approved the study. Caregivers gave written consent; when possible, children 7 years and above assented. A diagnosis of ASD was confirmed prior to participation by a licensed clinician at CADB using the Autism Diagnostic Observation Schedule (ADOS-2, Modules 13) or the Adapted ADOS Module 1. CSS for SA and RRBs were calculated. IQ scores were calculated from developmentally appropriate cognitive testing [16].

The target-child and their caregiver completed either a 1-week or an 8-week study that involved coming to the CADB clinic on one occasion (1-week protocol) or three separate occasions (8-week protocol) and completing study procedures in their home. Briefly, during the first clinic visit, caregivers were trained on operating the LENA DLP. The target-child wore a t-shirt that contained a pocket for the LENA DLP located on the chest during data collection. All participants completed recordings with the LENA device placed in the t-shirt. In the clinic, the target-child completed a series of standardized assessments while wearing the LENA device for $\tilde{5}0$ min. The assessments included a modified version of the Brief Observation of Social Communication Change [17], as well as the Purdue Pegboard task for 10 min, playing a puzzle game on an iPad for 10 min and watching a series of Pixar short movies on the iPad for 10 min. In the home, caregivers were instructed to record their childs speech for 3 days a week up to 1.5 hr per day during week 1 (for both the 1-week and 8-week protocol), as well as weeks 4 and 8 (8-week protocol only). They were encouraged to record their childs speech during times when the child would likely to be talking with them (e.g., dinner time).

2.2.2 Infants Dataset: IBIS Study

Infant Brain Imaging Study (IBIS) is a large and unique prospective longitudinal ongoing study of the Autism Centers of Excellence (ACE) Network, funded by the National Institute of Health, USA. This study was started in 2007 and has so far enrolled $\tilde{7}00$ infant siblings at high familial risk for autism. The data has been collected through two separate waves (IBIS-1 between 2007-2012 and IBIS-2 from 2012-2017). The third wave (IBIS-SA) which began in 2018 focuses on following these children into school age. Clinical data collection occurs at four sites the University of North Carolina (UNC), Childrens Hospital of Philadelphia, University of Washington (UW, Seattle), and Washington University in St. Louis (WUSTL). Data from all sites are maintained by a Data Coordination Center at the Montreal Neurological Institute. The study includes both neuroimaging and behavioral assessments of high-risk infant siblings and low risk infants at different time points from birth through 3 years of age and later into school age.

Among many modalities of data collected for children in the IBIS study, one was audio collected using the LENA device in the child's natural home environment at 9 month and 15 month of ages respectively. Audio data of a small subset of children available in the study was annotated using the GUI based annotation toolkit developed. This small subset of children is the main object of study in this work with regard to the IBIS study. We term this dataset as the *infants dataset*.

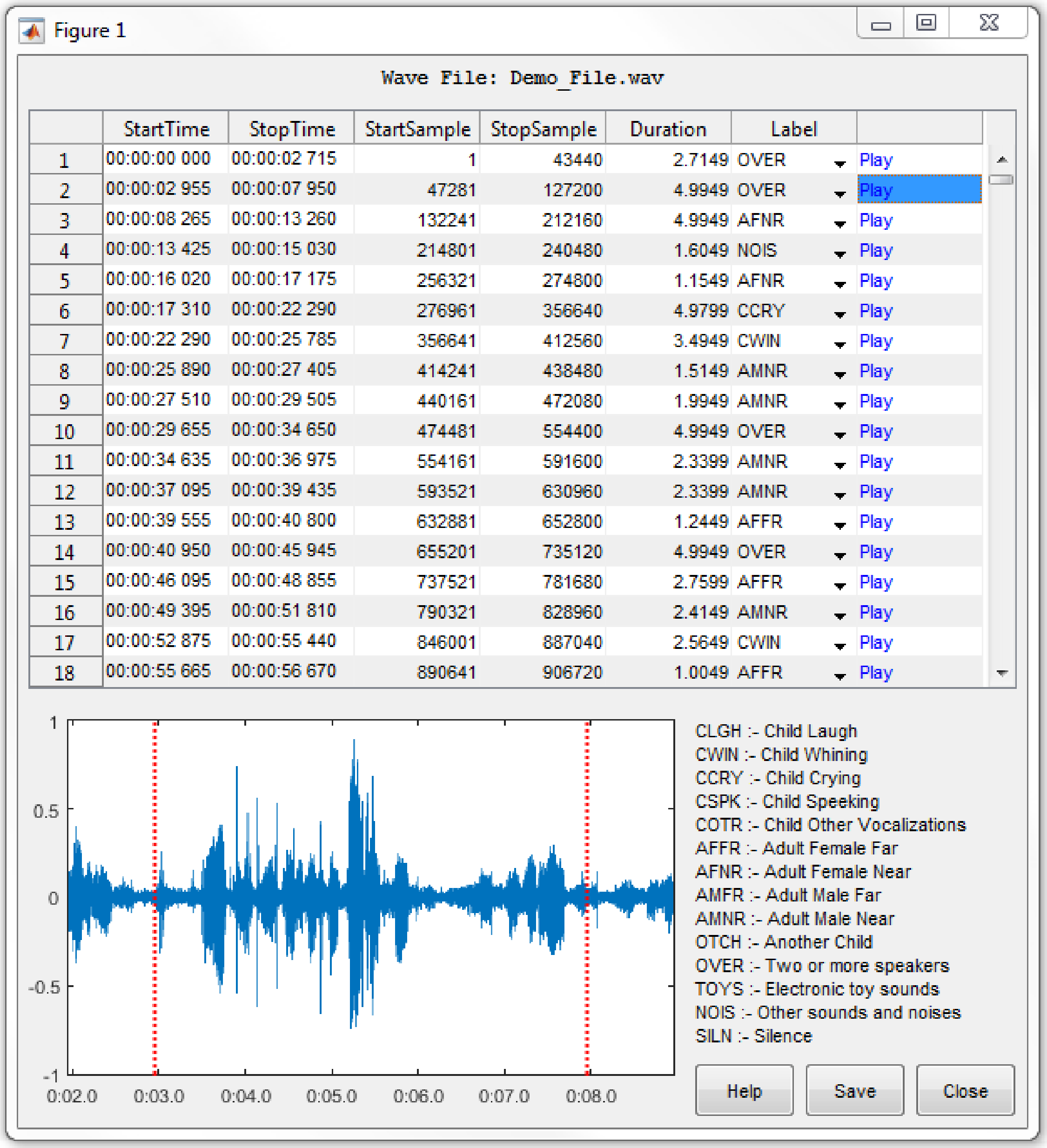
2.3 GUI Based Annotation Toolkit

Since, existing tools for annotating this data were deemed too slow for our limited application, a more efficient system was employed. A GUI based annotation toolkit was developed, which provided audio segments for listeners to annotate based on the segmentation of LENA system. The segments were presented to human listeners who were asked to label them into one of the following broad categories: a) child vocalization, b) adult vocalization, c) silence, d) environmental noise or e) multiple speakers. We also allowed finer-grained labels of child vocalizations which included child laughing, child whining, child crying, or child speaking.

A GUI based annotation software was developed by us to label audio segments based on LENA segmentation. The software takes the LENA raw audio files and CSV files generated by ADEX software as the input and generates a CSV file which contains human annotated labels. This software can be used to label very small portion of audio to train speaker models for a completely new LENA recording. It can also be used to label a completely new dataset for a new study in which we do have have any human annotated labels. The amount of human annotation required for a particular study depends upon the goals and requirements of the study. For most studies, a small subset of all data would be labeled for training and testing purposes.

Features and Requirements of annotation software

- The software is designed for Windows based machines.



Wave File: Demo_File.wav

	StartTime	StopTime	StartSample	StopSample	Duration	Label	
1	00:00:00 000	00:00:02 715	1	43440	2.7149	OVER	Play
2	00:00:02 955	00:00:07 950	47281	127200	4.9949	OVER	Play
3	00:00:08 265	00:00:13 260	132241	212160	4.9949	AFNR	Play
4	00:00:13 425	00:00:15 030	214801	240480	1.6049	NOIS	Play
5	00:00:16 020	00:00:17 175	256321	274800	1.1549	AFNR	Play
6	00:00:17 310	00:00:22 290	276961	356640	4.9799	CCRY	Play
7	00:00:22 290	00:00:25 785	356641	412560	3.4949	CWIN	Play
8	00:00:25 890	00:00:27 405	414241	438480	1.5149	AMNR	Play
9	00:00:27 510	00:00:29 505	440161	472080	1.9949	AMNR	Play
10	00:00:29 655	00:00:34 650	474481	554400	4.9949	OVER	Play
11	00:00:34 635	00:00:36 975	554161	591600	2.3399	AMNR	Play
12	00:00:37 095	00:00:39 435	593521	630960	2.3399	AMNR	Play
13	00:00:39 555	00:00:40 800	632881	652800	1.2449	AFFR	Play
14	00:00:40 950	00:00:45 945	655201	735120	4.9949	OVER	Play
15	00:00:46 095	00:00:48 855	737521	781680	2.7599	AFFR	Play
16	00:00:49 395	00:00:51 810	790321	828960	2.4149	AMNR	Play
17	00:00:52 875	00:00:55 440	846001	887040	2.5649	CWIN	Play
18	00:00:55 665	00:00:56 670	890641	906720	1.0049	AFFR	Play

CLGH :- Child Laugh
CWIN :- Child Whining
CCRY :- Child Crying
CSPK :- Child Speeking
COTR :- Child Other Vocalizations
AFFR :- Adult Female Far
AFNR :- Adult Female Near
AMFR :- Adult Male Far
AMNR :- Adult Male Near
OTCH :- Another Child
OVER :- Two or more speakers
TOYS :- Electronic toy sounds
NOIS :- Other sounds and noises
SILN :- Silence

Figure 2.1: Annotation Software Interface

Table 2.1: Comparison of LENA Annotation with Manual Annotation for Older Children Dataset

			LENA	Labels		
		Adult	Child	OVER	Other	OTCH
	Adult	10490	549	3486	3053	1263
Ground	Child	6400	10629	2124	1314	2557
Truth	OVER	3903	981	6217	965	591
Labels	Other	1505	1779	4765	43744	1527
	OTCH	212	285	439	332	1217

Table 2.2: Comparison of LENA Annotation with Manual Annotation for Infants Dataset

			LENA	Labels		
		Child	Adult	OVER	Other	OTCH
	Child	1396	54	587	111	156
Ground	Adult	105	684	291	96	80
Truth	OVER	474	424	1036	42	133
Labels	Other	135	57	285	460	67
	OTCH	84	35	122	39	202

- Its an easy GUI based application, in which the user clicks on a button to play audio and selects a label from a drop down list.
- The software can be easily modified to include labels according to the requirements of a study.
- The software only requires that MATLAB runtime version 8.4 (R2014b) is installed on the deployment machine.

2.4 Accuracy of LENA Software

In this section, accuracy of LENA software is evaluated for older children and infants datasets.

Table 2.3: Individual Child DetectorPrecision and Recall Statistics for 36 Subjects

Child Id	Recall	Precision
SANCOS	40.11	90.8
CAMPJU	65.9	86.36
NECHJE	33.09	60.0
SIMMJA	60.07	73.39
STERRA	27.77	70.18
DONOMA	32.65	64.0
TARARA	64.37	80.91
KOTBAN	57.79	72.89
NASIMA	46.76	55.86
SCOXED	43.89	72.48
HJIASO	36.33	80.4
MCDOLU	49.39	67.33
HJIATO	56.51	81.28
RIKOGA	43.21	64.71
KESMAB	44.21	72.5
JABLNI	3.41	17.07
HUGGOR	61.0	65.92
RIKOSA	28.96	67.26
BELMPE	0.0	0.0
ACEVNA	14.91	72.6
PEYSO	41.92	65.98
ARNOBO	13.58	60.42
KATXEL	24.47	46.29
WRAXSY	49.34	82.43
PELLDA	37.89	62.74
TENOJA	45.03	85.05
ZAGARU	6.83	38.0
BEHPAN	62.36	80.11
WEXXYU	21.24	49.55
SIMXMO	49.28	69.44
BAIXLU	71.59	86.54
MCGRFI	53.73	83.53
LEWICA	46.45	90.31
GRIFSP	71.39	92.31
STEIJO	39.73	72.95
KANDLI	54.73	62.7

2.4.1 Accuracy of LENA software for Older Children dataset

The LENA child detectors were trained using speech from children aged 12 months to 48 months. Here, we analyze if the LENA algorithms are accurate enough for older children aged 5 to 17 years old.

Table 2.1 shows the confusion matrix of LENA labels versus human annotation. The manually annotated labels in the table are deemed to be the ground truth. Two trained research assistants used the GUI based annotation toolkit to label roughly 1500 audio segments from clinic conditions and another 1500 from home conditions for every subject's audio recording."Child" and "Adult" are self-explanatory and mean audio segments which had child and adult utterances respectively. Label "OVER" includes segments which were marked as multiple speakers and overlap by human annotators. The LENA software label "OVER" denotes audio segments in which there is an actual overlap between two speakers. Note, there is no LENA label which corresponds to "multiple speakers. Label "Other" includes TV noise, environmental noise and silence. Label "OTCH" refers to another child. The overall recall and precision for child detection across the 36 subjects was 46.16% and 74.73% respectively, while the overall recall and precision for adult detection was 55.67% and 46.6 % respectively. As noted, label "OVER" included segments in which adult and child speakers spoke one after the other and actual overlap during which adult and child speakers were speaking at the same time (cross-talk). The situation when adult and child speakers spoke one after the other was an anomaly of segmentation software. However, only 11.47% of all utterances were labeled as "OVER." The recall for LENAs child detector was 46.16% which means more than half of all child utterances were missed from a detection point of view. Among the child segments that were missed, the most common confusion was child utterances being misclassified as adult utterances (roughly 52% of all child utterance mis-classifications were this). The LENA detector models often confuse older children (aged 5-17 years) as being adults. This behavior is not surprising since LENA child models were trained using children that were younger than those in the present study. The Child detector precision of 74.73% makes it useful in calculating certain statistics. Since we are 75% sure that a detected child utterance is really a child, other observations can be made, such as voice characteristics or conversational turn likelihoods. However, this relatively higher precision comes at the cost of low recall, which make it difficult to estimate absolute counts, such as number of utterances or number of conversational turns. As noted earlier, a significant number of child utterances are misclassified as adult utterances (roughly 28%). This observation further makes useful computation of conversation-based language behavior statistics difficult. The adult detector precision (46.6%) and recall (55.67%) values are not high enough to be able to compute accurate conversation based language behavior statistics. A high precision value could have been useful in obtaining some parts of audio in which adult is surely speaking. This in turn can then be used to analyze child's response to the obtained adult speech. For this study, the presence of another child in the audio was very low (2.25% of all segments). However, its important to note that precision for detecting another child was 17.01%. One reason for such a low precision value was that only 11.1% of all primary child utterances were classified as another child. This was the second most common error among all child utterance misclassification.

Table 2.3 shows the individual recall and precision of child detectors across all the 36 subjects. The recall for child detection across all 36 subjects was low. The maximum recall observed across all the subjects was 71.59 %. However, only 7 out of 36 subjects had a recall rate of greater than 60%, which suggests that the LENA child detectors miss a lot of audio segments that have child utterances. The maximum precision observed across all the subjects was 92.3% which is very accurate, however 24 subjects out of the 36 subjects analyzed had a precision rate less than 80%.

Table 2.4: Individual Child Detector Precision and Recall Statistics for 12 Subjects

Child Id	Recall	Precision
1	41.96	58.02
2	49.1	81.95
3	47.55	65.97
4	68.53	52.69
5	58.45	78.28
6	47.86	69.07
7	58.65	30.35
8	49.62	65.02
9	38.9	95.35
10	35.37	33.33
11	45.95	85.0
12	56.07	47.62

2.4.2 Accuracy of LENA software for Infants dataset

Experiments performed on the older children dataset suggests that the LENA software is not accurate enough for meaningful analysis for older children age group. Since, the LENA software was designed using acoustic data from children aged 12 months - 48 months, it is expected that LENA acoustic models do not generalize to children of other age groups. In this section, the accuracy of LENA software for infants dataset described before is analyzed. The acoustic data for children in this dataset was collected either at 9 month stage of their development or at the 15 month stage of their development.

The LENA audio data for the 12 infants present in the infants dataset was annotated by 2 annotators using the GUI annotation toolkit described before. In order to compute the accuracy of LENA software compared to human annotators, only audio segments which both the annotators agreed were considered for analysis.

Table 2.4 shows the individual child detector precision and recall statistics for the 12 infants present in the infants dataset. The maximum recall of child segments observed over all infants was only 68.53%. The average recall rate over all 12 infants was only 49.83%, meaning on an average the LENA child detector only detected about 50% of true child segments. The precision rate observed across 12 infants showed a lot of variation. For Child ID's 2,5,9, and 11 the precision rate observed was about 80% or more. However, on the other spectrum, the precision rate observed for Child Id's 4,7,10, and 12 the precision rate observed was around 50% or less. This shows, that the child detector precision values are not consistent across all the subjects. The average precision rate observed across 12 infants was 63.55%, which is too low for meaningful analysis.

2.5 Summary and Discussion

The comparison of LENA software with human annotation for both older children dataset and infants dataset suggests that the accuracy of LENA detectors is not high enough. There are possibly many different factors contributing to this phenomenon. One contributing factor is probably the fact that the age group (5-17 years) of the children present in the older children dataset is significantly different to the age group (12 months - 48 months) on which the LENA software was trained and was probably intended to have been used for. Even the children present in the infants dataset (9 months) either do not fall in the age group on which the LENA software was trained or are at the extreme end (15 months) of the age group on which the LENA software was trained. The audio data corresponding to 8 of the 12 infants present in the infants dataset (Child ID's 1,4,5,6,7,8,9, and 10) was collected at 15 months. The LENA child detectors as seen from table 2.4 were very inaccurate for this group of 8 infants (Average Recall Rate: 49.9%, Average Precision Rate: 60.23%) and in fact, show no improvement compared to the group of 4 infants whose audio data was collected at 9 month stage. (Average Recall Rate: 49.67%, Average Precision Rate: 70.13%). This observation, even though the number of children present in the analysis is low suggests that the LENA child detector does not do well at the lower extreme end of age group (near 12 months) on which it was trained. Another probable contributing factor is that LENA detection models use variation of Gaussian Mixture Model (GMM) based methods, which perform significantly worse than current state of the art methods, particularly i-vector methods on most speaker identification and diarization tasks. In addition to the two possible reasons mentioned, another major contributing factor is that LENA speaker detector models are completely subject independent. Ideally, it is desirable to have highly accurate subject independent speaker detector methods, as they would not require any new annotation for training. However, if the subject independent models are not accurate enough for meaningful analysis, development of methods which do some form of speaker dependent training is required. In this work, semi-supervised i-vector methods and unsupervised i-vector methods are developed, which are shown to be accurate for the two datasets analyzed in this work. The first method called the semi-supervised i-vector based diarization uses some small amount of data (2 minutes of speech) to train speaker specific models present in the audio recording, while the second method called the unsupervised i-vector based diarization uses no labeled data to diarize a given audio recording. The development and analysis of these methods is described in chapter 3.

CHAPTER 3

AUDIO DIARIZATION FOR LENA DATA

As was described in chapter 2, the accuracy of LENA acoustic detectors is not good enough for meaningful analysis and development of applications such as design of objective measures to track treatment methods, early diagnosis of autism spectrum disorder related symptoms etc. In order to do accurate audio diarization, i-vector based diarization methods were developed for two scenarios: 1) Some amount of labeled data is available for all the speakers present in the given LENA audio recording and 2) No labeled data is available for any of the speakers present in the given LENA audio recording. The methods developed for the first scenario are termed as semi-supervised i-vector based diarization, while the methods developed for the second scenario are termed as unsupervised i-vector based diarization methods in this work.

This chapter describes a) general theory and practice of i-vector based methods b) development of semi-supervised methods for older children and infants dataset, and c) development of unsupervised i-vector based diarization methods for older children and infants dataset.

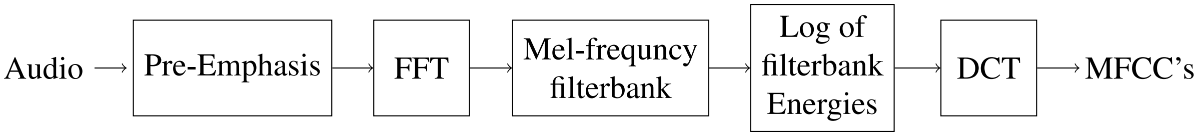


Figure 3.1: Steps in MFCC feature extraction

3.1 MFCC Feature Extraction and Universal Background Models

In order to understand i-vector based diarization methods developed in this work, it is essential to first know the basics of general feature extraction pipeline followed in speech and audio processing tasks.

Any digital audio recording is just a stream of real-valued numbers encoded using some encoding standard with some precision sampled at some sampling rate (eg. LENA audio is encoded as signed integer PCM with 16 bit precision at 16 kHz sampling rate). So, any audio utterance or segment is then just a set of numbers. In most speech and audio processing tasks, the initial feature extraction of a speech utterance or segment involves dividing the audio utterance into overlapping intervals of fixed size called as "frames". Typically, these frames are of 20-30 ms duration with an overlap of 10-15 ms. In this work, 20 ms frames (a set of 320 numbers) with 10 ms overlap were used. The raw digital audio per frame is passed through a window function, generally hamming window [18]. Then, some fixed dimensional acoustic feature vector is extracted per frame.

3.1.1 MFCC Coefficients

In speaker recognition and diarization, mel-frequency cepstral coefficients (MFCC's), along with their deltas, and delta-deltas are the typical choice of feature representation per frame.

Figure 3.1 shows the basic steps involved in the computation of MFCC features. After passing the audio data through the windowing function (typically Hamming), the higher frequencies of the data are amplified via a linear “pre-emphasis” filter and the discrete spectrum is computed using the fast Fourier transform (FFT). The spectral representation of the audio data in the form of FFT is passed through a pyschoacoustically motivated mel-frequency filterbank. Each of the filters in the filterbank is triangular and computes the energy average around the center frequency of each triangle. The center frequencies of the filterbank are linearly spaced on the mel-frequency scale, which was designed to approximate the behavior of the human auditory system. Finally, the discrete cosine transform (DCT) is used to reduce the correlation between the filters. For speaker recognition and diarization applications, typically coefficients 1-19 as well as the log of the energy of the audio signal is used, giving a 20 dimensional vector. In order to incorporate some temporal information to the features extracted, estimates of the first-order and second-order temporal derivatives are obtained, known as delta and delta-delta coefficients, respectively. Thus, a 60 dimensional acoustic feature vector is obtained per frame of the audio data.

3.1.2 Universal Background Models

Before the advent of i-vector based methods, the most popular technique for speaker recognition and diarization problems was based on Gaussian Mixture Models (GMM’s), introduced in [19–21]. An important notion in GMM based speaker identification and diarization approaches is that of Universal Background Models, generally referred to as UBMs. The front-end of I-vector based methods, as would be shown later are build on top of UBMs.

A Universal Background model is a generative speaker-independent model which is used to capture the variability encountered in the frame-wise acoustic feature vectors. It is modeled using a GMM with a large number of mixture components. For a given GMM, let C be the number of components in the mixture. Let, x be an F-dimensional feature vector, and μ_i and Σ_i be the mean and covariance matrix of the i^{th} mixture component of sizes F

and $F \times F$ respectively. Let, w_i represent the weight of the i^{th} mixture component. Suppose, $\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{i=C}$ represents the parameter set of the GMM. Then, the probability of x given λ is

$$P(x|\lambda) = \sum_{i=1}^{i=C} w_i \mathcal{N}(x; \mu_i, \Sigma_i) \tag{3.1}$$

where,

$$\mathcal{N}(x; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{F}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \tag{3.2}$$

which is the standard multi-dimensional gaussian.

In an ideal sense, a UBM can be interpreted as the background model for all speakers. That is, this model should capture the acoustic feature variation among all the speakers that it is going to encounter. Thus, in order to achieve high accuracy on speaker recognition and diarization tasks for both traditional GMM-UBM approaches and i-vector based approaches, it is essential that the the distribution of acoustic feature vectors used to train the UBM model is similar to the distribution of acoustic feature vectors during testing.

The parameter set λ of the UBM is trained using an Expectation-Maximization (EM) algorithm [19–22]. Suppose, $X = \{x_1, x_2, x_3, x_4, ..., x_N\}$ is the set of acoustic feature vectors from all the speakers available for training the UBM. Then, the log-likelihood of the acoustic feature vectors for a parameter set λ is given by

$$\log(P(X|\lambda) = \sum_{i=1}^{N} P(x_i|\lambda) \tag{3.3}$$

During the EM algorithm, it is guaranteed that the log likelihood described by equation 3.3 increases after every iteration. Let, $\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{i=C}$ be the current value of the parameter set of the UBM at during EM training. Let, $\lambda' = \{w_i', \mu_i', \Sigma_i'\}_{i=1}^{i=C}$ be the updated values of the parameter set of the UBM after one iteration of EM training. Then, the updated values of the parameter set are given as

$$w_i' = \frac{1}{N}\sum_{t=1}^{N} P(i|x_t, \lambda) \tag{3.4}$$

$$\mu_i' = \frac{\sum_{t=1}^{N} P(i|x_t, \lambda)x_t}{\sum_{t=1}^{N} P(i|x_t, \lambda)} \tag{3.5}$$

$$\Sigma_i' = \frac{\sum_{t=1}^{N} P(i|x_t, \lambda)(x_t - \mu')(x_t - \mu')^T}{\sum_{t=1}^{T} P(i|x_t, \lambda)} \tag{3.6}$$

An important thing to note is that UBM training does not require any labeling information. In practice, the number of mixture components chosen is a large number, generally 1024 or 2048. There are no explicit reasons for choosing exactly these numbers, it has just become a matter of tradition in the speaker recognition and diarization community. Any number of the mixture components of the same order should work equally well.

In traditional GMM-UBM based approaches for speaker recognition and diarization, the speaker model for a specific speaker is again modeled using a GMM. Since, the amount of data available per speaker might be less, a maximum aposteriori (MAP) adaptation is done to adapt the specific speaker model to the UBM [21]. Generally, only the means of the UBM are adapted. Let s be some specific speaker for whom a speaker detector is to be designed. Let, GMM_s be the speaker model for s after MAP-adapting the UBM parameters to s using the training data for speaker s. Then, the goal is to determine if speaker s is present in some audio utterance Y. Let, audio utterance Y be a set of acoustic feature vectors $Y = \{y_1, y_2, y_3, ..., y_M\}$ corresponding to M frames. Then, the log-likelihood that Y belongs to from GMM_S compared to UBM is evaluated. Let, this be represented by

$L(Y)$. It is given by,

$$L(Y) = \log \frac{P(Y|GMM_s)}{P(Y|UBM)} = \sum_{t=1}^{t=M} \log P(y_t|GMM_s) - \log P(y_t|UBM) \qquad (3.7)$$

Then, $L(Y)$ is tested against some pre-trained threshold θ. If, $L(Y) > \theta$, then Y is said to have been uttered by speaker s.

In i-vector based speaker recognition and diarization approaches, an important step during the front-end training, as will be seen in section 3.2 is training of a Universal Background Model. Section 3.2 will describe the basic theory of i-vector based methods. The semi-supervised and unsupervised methods developed for LENA audio data, and described in sections 3.3 and 3.4 are built on i-vector methods.

3.2 Basic theory of i-vector Methods

Until very recently, i-vector based methods with probabilistic linear discriminant analysis (PLDA) as back-end were the state of the art methods [23–26] for speaker recognition and diarization tasks. These methods have largely been deployed in traditionally well studied domains of application like telephony, broadcast news, and meeting room data. These domains have large curated datasets with significant amount of labeled data available. The US National Institute of Standards and Technology (NIST) conducts a speaker recognition evaluation (SRE) periodically to enable progress for speaker recognition research [27–29]. The domain mainly studied in SRE evaluations is telephony speech with conversational telephone speech being the focus of the current SRE 2019 evaluation [29].

The front-end of i-vector based methods consist of representing an audio utterance into a fixed-dimensional vector which captures the speaker and channel characteristics of the audio utterance. The important thing to note is that audio utterances of different durations are mapped to vectors of the same dimension.

3.2.1 I-vector Modeling and Training

Let, S be a speaker space, where points in S represent different speakers. The assumption in most Speaker ID techniques is that S is a low dimensional manifold embedded in some higher dimensional space [20, 21, 30–33]. Usually, this higher dimensional space is supervector space, in which, a supervector for some speaker s is obtained by adapting a Universal Background Model (UBM) for s.

In most previous techniques, such as Joint Factor Analysis (JFA), a low dimensional speaker representation is computed directly from supervector space i.e channel compensation is assumed directly into the model. However, in the i-vector model, a low dimensional representation of an audio utterance is assumed which has both speaker and channel characteristics. Channel compensation is performed at the back-end using techniques such as Linear Discriminant Analysis (LDA), Probabilistic Linear Discriminant Analysis (PLDA) [24].

A Universal Background Model (UBM) is trained to model variability in acoustic feature vectors computed from acoustic data comprising of many different speakers. For an audio utterance, mel-frequency cepstral coefficients (MFCC) features along with deltas and delta-deltas are computed for every acoustic frame. A Gaussian Mixture Model (GMM) is trained as the UBM.

Terminology and Setup

Let, U be a Universal Background Model modeled as a F dimensional Gaussian Mixture Model (GMM) with C mixture components.

Supervector: It refers to the CF dimensional vector obtained by concatenating the F-dimensional mean vectors corresponding to an utterance GMM.

The assumption, in almost all speaker ID techniques is that, the utterance supervectors lie in a low dimensional space and hence, can be represented by small number of parameters i.e if m and B represent the mean and covariance matrix of supervectors of utterances, then B is of low rank. For the i-vector method, the speaker and channel dependent supervector M is modeled as

$$M = m + Tw \tag{3.8}$$

where m is the speaker and channel independent supervector, T is a rectangular matrix of low rank R, and w is a random vector with standard normal distribution $\mathcal{N}(0, I)$. Therefore, M is normally distributed with mean m and covariance matrix TT^*.

Let $M(u)$ represent a supervector associated with an utterance u. For each mixture component c, define $M_c(u)$ as the subvector of $M(u)$ which corresponds to it. It is assumed that there is a covariance matrix Σ_c, such that for any utterance u, acoustic observation vectors associated with the mixture component are normally distributed with mean $M_c(u)$ and covariance matrix Σ_c. Let Σ denote the $CF \times CF$ block diagonal matrix whose diagonal blocks are Σ_1, Σ_2, Σ_3, ..., Σ_C.

Let $X(u)$ be the acoustic feature vectors associated with utterance u. Baum-Welch statistics are extracted from $X(u)$.

$$\gamma_t(c) = P(c|X_t, UBM) \tag{3.9}$$

Then, let

$$N_c(u) = \sum_{t=1}^{L} \gamma_t(c) \tag{3.10}$$

$$F_c(u) = \sum_{t=1}^{L} \gamma_t(c)(X_t - \mu_c) \tag{3.11}$$

$$S_c(u) = \sum_{t=1}^{L} \gamma_t(c)(X_t - \mu_c)(X_t - \mu_c)^* \tag{3.12}$$

Let, $N(u)$ be the $CF \times CF$ block diagonal matrix whose diagonal blocks are $N_1(u)I$,$N_2(u)I$,...,$N_C(u)I$, where I is the $F \times F$ identity matrix. Let $F(u)$ be the $CF \times 1$ vector obtained by concatenating $F_1(u)$,$F_2(u)$,...,$F_C(u)$. Let, $S(u)$ be the $CF\times$ block diagonal matrix whose diagonal blocks are $S_1(u)$, $S_2(u)$,...,$S_C(u)$.

The new vectors w are a low dimensional representation of M and are referred as i-vectors.

EM Steps

The main computation that needs to be done for estimating T and Σ and for MAP adaptation for utterance u is to calculate the posterior distribution of $w(u)$ given utterance data.

For each utterance, let $l(u)$ be a $R \times R$ matrix given as

$$l(u) = I + T^{*}\Sigma^{-1}N(u)T$$

Then, for each utterance u, the posterior distribution of $w(u)$ given $X(u)$, T, and Σ is Gaussian with mean

$l^{-1}(u)T^{*}\Sigma^{-1}F(u)$ and covariance matrix $l^{-1}(u)$.

The E steps consists of computing $E[w(u)]$ and $E[w(u)w(u)*]$ given current estimates of T and Σ. This has a nice analytic solution.

In the M step, let T and Σ be new estimates of the model parameters. Then, T is the solution of

$$\sum_{u} N(u)TE[w(u)w(u)^{*}] = \sum_{u} F(u)E[w^{*}(u)] \tag{3.13}$$

and, for each component c,

$$\Sigma_c = \frac{1}{n_c}(\sum_{u} S_c(u) - Z_c) \tag{3.14}$$

where $n_c = \sum_c N_c(u)$.

The new estimates of T and Σ are guaranteed to increase the log likelihood of all data. The details of computation can be found in [30].

Once, T and Σ are trained, then for any utterance u, the mean of the posterior distribution of $w(u)$, given acoustic frames corresponding to$u$ is computed. This mean is termed as the i-vector.

3.2.2 PLDA

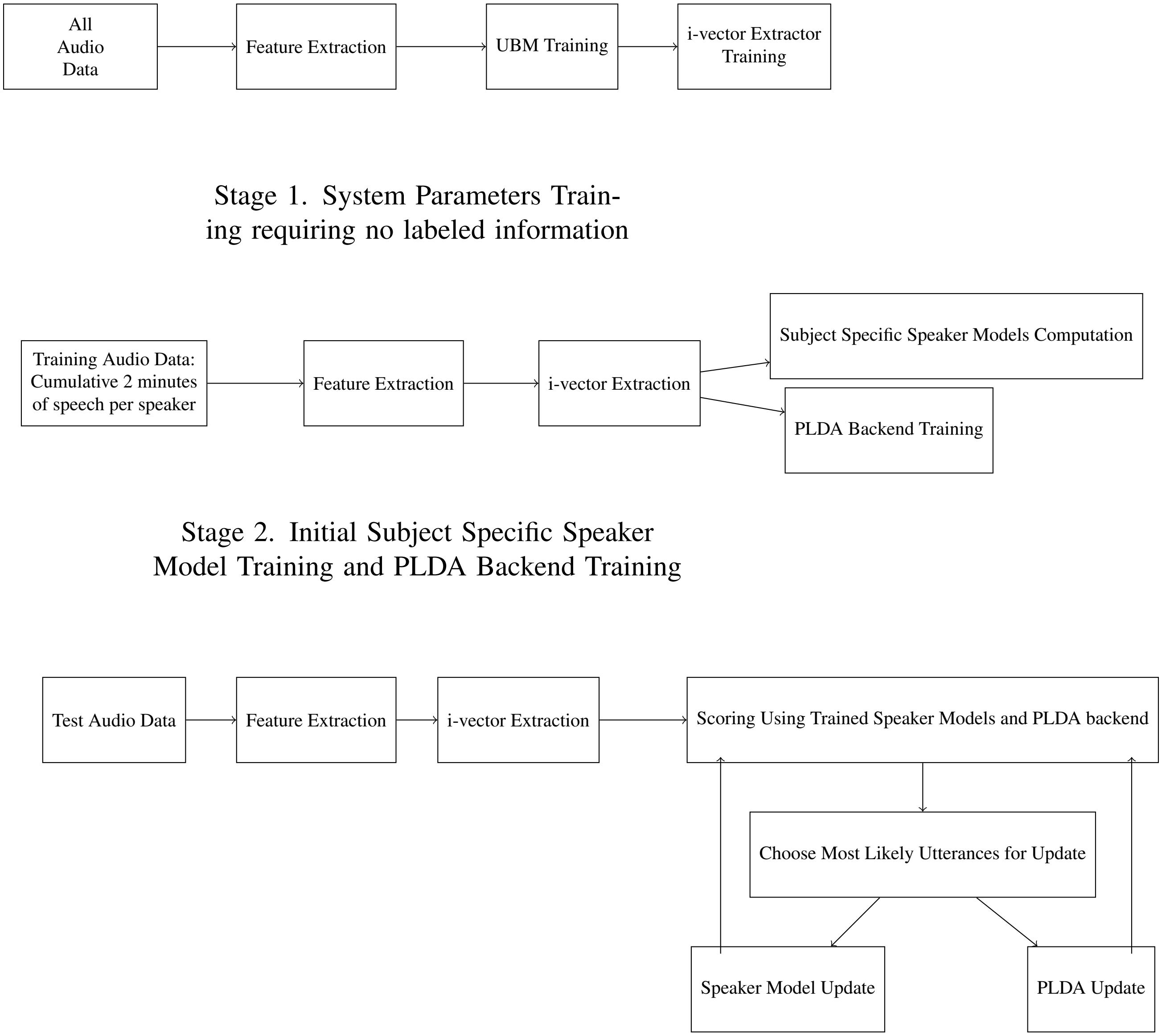
Probabilistic linear discriminant analysis (PLDA) is the back-end which is typically used to compute a likelihood if two i-vectors belong to the same speaker in speaker recognition and diarization systems.

In PLDA scheme, it is assumed that the i-vector $w(u)$ corresponding to utterance u can be modeled as

$$w(u) = \mu + Vx + \epsilon(u) \tag{3.15}$$

Here, $\mu + Vx$ can be viewed as the speaker-specific part which captures the identity of the speaker and does not depend on the particular utterance u and $\epsilon(u)$ can be viewed as the utterance dependent channel component. $\epsilon(u)$ is assumed to be gaussian with zero mean and a covariance matrix W. μ is a global parameter independent of the mean, while V is the speaker subspace model and x is a latent speaker identity vector that has a standard gaussian prior. The parameters $\{\mu, V, W\}$ are trained using an EM algorithm. The details of EM training and computation can be found in [24].

Once, $\{\mu, V, W\}$ are trained, for any two i-vectors w_1 and w_2, the likelihood that the two i-vectors share the same latent speaker identity is computed.



Stage 1. System Parameters Training requiring no labeled information

Stage 2. Initial Subject Specific Speaker Model Training and PLDA Backend Training

Stage 3. Speaker Model and PLDA update

Figure 3.2: Semi Supervised I-vector based Annotation System

3.3 Semi-Supervised I-vector based Diarization

The LENA audio data which is the focus area of this work is more challenging compared to domains like telephony, broadcast news, and meeting room data. There are three main challenges compared to the traditional studied domains:

- LENA audio is very noisy as the microphone is attached to the clothing of child.
- LENA audio is single-channel data.
- The amount of labeled data for training models is very less compared to traditional studied domains in speaker recognition and diarization tasks.

Compared to training speaker recognition and diarization systems for traditional systems, there are fewer unique speakers present in the datasets that are studied here. However, there is lot of speech data present for all speakers in LENA recordings. In the semi-supervised method, this property for this dataset is exploited. These methods are developed for the scenario in which some amount of labeled data is present for every speaker present in the audio recording. Initially using the limited amount of labeled data (here 2 minutes of speech per speaker), an initial subject specific i-vector model is constructed. Using these initial models and an initial estimate of the PLDA model, unlabeled audio utterances are tested against these models and then these unlabeled audio utterances are used to update the subject-specific speaker models.

The semi-supervised algorithm is illustrated in figure 3.2, in the first stage, i-vector extractor training is done which does not use any labeled data. In the second stage, 2 minutes of speech is used to compute initial speaker models. In the third stage, data from unlabeled audio is used to update the subject-specific speaker models.

Algorithm 1 Semi-Supervised I-vector Training Algorithm

```
Initial Training: 1) Train UBM and I-vector extractor (T) matrix 2) Use labeled information to train initial i-vector speaker models $S_{models}$ and PLDA scorer

for fixed number of iterations do
    while parse through unlabeled audio segments do
        Choose audio segments for update using current $S_{models}$ and PLDA
    end while
    Update $S_{models}$ and PLDA
end for
```

3.3.1 System Development and Data Analysis

In this section, we describe the development of our semi-supervised i-vector based diarization system. To this end, we will show that it is possible to detect child, adult female, and adult male segments with high accuracy for a LENA recording. We term the developed system "semi-supervised" because it uses unlabeled data to update or retrain speaker and PLDA models. Figure 3.2 illustrates the various stages involved in system development.

We trained i-vector based vocalization detectors for clinic and home environments. The LENA audio data was single channel data sampled at 16 kHz. We used Kaldi toolkit [34] to run our experiments. A total of 20 MFCC features plus deltas and delta deltas were extracted based on 20 milliseconds (ms) Hamming windowed data with a frame rate of 100/second. A basic energy based voice activity detection (VAD) system was used to select voice only frames.

Stage 1: Parameter Training Using No Labeled Information

A 1024 Gaussian Mixture Model (GMM) using around 20 hours of data was trained as a Universal Background Model (UBM). An i-vector extractor was trained to generate a 400 dimensional i-vector for a given audio segment. Therefore, every audio segment was mapped into a 400 dimensional i-vector. No labeling information was required to train UBM and i-vector extractor.

Table 3.1: System Parameter Values Used in the i-vector system

Parameter Name	Value
Acoustic Features	20 dimensional MFCC's
Frame Length	20 ms (Hamming Window Used)
Frame Rate	100 frames/sec (Frame Shift 10 ms)
Universal Background Model	1024 Gaussians
i-vector dimensionality	400

Stage 2: Initial Speaker and PLDA Models

The initial subject specific speaker models for every child and adult were computed using 2 minutes of speech per speaker. To be precise, for any given speaker a set of audio segments was chosen such that the cumulative amount of speech after being processed though VAD in these segments was approximately 2 minutes. The mean of the i-vectors corresponding to these segments was deemed to be the speaker model for that speaker. For child speakers, separate models were formed in clinic and home conditions.

The data used to train speaker models i.e., 2 minutes of speech per speaker, was used to train the initial PLDA back-end scorer. We used additional data of 2 minutes comprised of toys, TV, environmental noise, silence from every recording. The underlying assumption specific to our data (due to the nature of annotation process) during the training of PLDA scorers is that all female labels in a given recording correspond to the same speaker. The same is assumed for child and male speakers. Another assumption is that speakers of one recording are distinct from speakers of another recording. However, all the interactions with the children in clinic were done by two research assistants. Therefore, at most, data from 2 recordings could have been used for training the PLDA. At the time of running the experiments, information about which research assistant interacted with a particular child was not available. Hence, training a PLDA scorer for a recording corresponding to a particular child used data from all child and adult speakers in the home environment, all child speakers from the clinic environment, and only one adult which interacted with the particular child.

3.3.2 Stage 3: Speaker and PLDA Models Update

We used unlabeled data to retrain our initial speaker and PLDA model updates. The i-vectors corresponding to unlabeled audio segments were tested against subject specific trained models trained in stage 2 using PLDA scorer trained in stage 2. The likelihood scores obtained were used to choose the most likely utterances for update. The speaker models and the PLDA back-end were then retrained using the limited labeled audio segments and the newly obtained most likely utterances.

Trials Generation

In order to test the accuracy of our i-vector based vocalization detectors, we created trials in which the speaker models were tested against i-vectors obtained from segments not used in training the speaker models and PLDA scorers. For example, let $Child\_1_{clinic}$ denote the speaker model of $Child\_1$ child in clinic conditions. Then, $Child\_1_{clinic}$ was tested against i-vectors corresponding to child segments, female segments, toy segments, noise segments, and silence segments from $Child\_1$'s LENA recording in clinic conditions, which were not used in any training. The PLDA scorers computed the likelihood that a given test i-vector belongs to $Child\_1_{clinic}$ speaker model.

An i-vector based diarization system with front end similar to our system, for LENA recordings of children aged 2.5 to 5 years, recorded in a childcare center was presented in [35] and had an accuracy of 69$\%$. Our system, however, used PLDA metric as the back-end for scoring, in contrast to [35], which used a Support Vector Machine (SVM) based scoring as the back-end. Another distinction was that our system does scoring against subject specific speaker models, obtained using two minutes of speech per speaker, while the system in [35] does subject independent supervised classification of primary child, secondary child (other children with which the primary child interacts at the childcare center) and adult vocalizations.

Table 3.2: Accuracy of semi-supervised i-vector based child detectors in clinic conditions for older children

Child Id	Number of Trials	EER
KATXEL	710	0.0
MCGRFI	636	11.0919
RIKOGA	725	12.2102
BEHPAN	685	8.3189
BAIXLU	646	8.2
GRIFSP	678	5.1836
DONOMA	672	11.6427
MCDOLU	622	11.0727
STERRA	553	12.1086
ACEVNA	639	16.3907
LEWICA	572	20.4819
HJIASO	684	14.8532
SANCOS	654	15.8182
JABLNI	514	7.5594
WEXXYU	634	12.2995
HJIATO	637	10.0694
WRAXSY	660	17.8218
TENOJA	708	28.9568
PELLDA	733	7.2327
CAMPJU*	739	15.24

3.3.3 Results for Older Children

The accuracy of the i-vector based vocalization detectors was tested using Equal Error Rate (EER)%. The equal error rate corresponds to the operating point of a detector at which the false rejection rate and false acceptance rate are equal. Tables 3.2 and 3.3 describe the accuracy of child and female vocalization detectors in clinic conditions. From table 3.7, the mean EER for child detectors in clinic conditions was 12.17%, which is highly accurate. The EER rates were less than 20% for almost all child speakers in the clinic conditions, which suggests that the methods are consistently accurate. The mean EER for female detectors in clinic conditions was 19.53% which is quite accurate, but not as accurate as female detectors in home conditions where the mean EER rate was 15.43%. The main reason we suspect is the fact that for training the PLDA scorer, two minutes of speech

Table 3.3: Accuracy of semi-supervised i-vector based female detectors in clinic conditions for older children

Child Id	Number of Trials	EER
BEHPAN	685	20.7071
SANCOS	654	21.0953
ACEVNA	639	13.7405
STERRA	553	28.3843
CAMPJU	632	19.4561
BAIXLU	646	24.1176
HJIATO	637	16.3534
MCGRFI	636	22.7477
WEXXYU	634	20.9924
GRIFSP	678	20.7436
MCDOLU	622	14.5455
DONOMA	672	11.1517
LEWICA	572	23.8482
HJIASO	684	22.7848
WRAXSY	660	13.933
JABLNI	514	25.1244
PELLDA	733	22.4961
RIKOGA	725	9.5159
TENOJA	708	32.8358
KATXEL	710	6.051

Table 3.4: Accuracy of semi-supervised i-vector based child detectors in home Conditions for older children

Child Id	Number of Trials	EER
PELLDA	411	8.3799
SANCOS	720	6.3618
BEHPAN	677	9.6931
HJIASO	482	12.8358
GRIFSP	620	5.1625
JABLNI	504	7.2261
LEWICA	628	10.8527
HJIATO	630	9.188
KATXEL	564	14.1717
BAIXLU	497	14.0212
WRAXSY	654	16.3551
STERRA	402	6.9638
RIKOGA	510	17.8571
MCGRFI	701	8.5455
ACEVNA	452	12.9477
MCDOLU	554	9.5785
TENOJA	534	17.4468
CAMPJU*	304	8.55
WEXXYU*	536	19.96
DONOMA*	267	4.98

Table 3.5: Accuracy of semi-supervised i-vector based female detectors in home Conditions for older children

Child Id	Number of Trials	EER
CAMPJU	210	9.0909
BEHPAN	677	17.5768
ACEVNA	452	15.427
HJIATO	630	11.9247
LEWICA	628	22.2846
MCGRFI	701	14.8021
JABLNI	504	21.9089
MCDOLU	554	10.4592
BAIXLU	497	9.6096
STERRA	402	19.0104
RIKOGA	510	12.3737
WRAXSY	654	11.7647
TENOJA	534	16.8016
HJIASO	482	13.1098
DONOMA	211	10.8911
WEXXYU	490	29.7921

Table 3.6: Accuracy of semi-supervised i-vector based male detectors in home Conditions for older children

Child Id	Number of Trials	EER
SANCOS	720	5.8333
STERRA	402	6.1972
GRIFSP	620	5.4054
KATXEL	564	8.4091
MCGRFI	701	7.0203

Table 3.7: Summary of Accuracy of semi-supervised i-vector based detectors for older children

Detector Type	Mean EER
Child Clinic	10.98
Female Clinic	16.64
Child Home	9.71
Female Home	12.35
Male Home	6.51

Table 3.8: Summary of Accuracy of semi-supervised i-vector based detectors for infants

Detector Type	Mean EER
Child Home	13.73
Female Home	18.97
Male Home	16.48

corresponding to female speakers was used from each recording in home conditions, while data from only one recording from clinic could be used. Another reason which might have negatively impacted the results could be that the time used to setup the interaction process with the child was also recorded and annotated. Multiple speakers spoke during the setup time but all were marked as the same label, which is adult female. The mean EER for child detectors in home conditions was 11.03%, which again is highly accurate. The EER rates for all child speakers was less than 20% for all child speakers, which again shows that the child detection was consistently accurate. The mean EER for male speakers in home conditions was 6.57%. The detectors were accurate in detecting both male and female speech, however the accuracy in males was much greater than females. The most likely reason for such behavior is that since the age of the children for our study was from five to fourteen years old, their voices were more similar to female voices than male voices.

3.3.4 Results for Infants

Similar to experiment conducted for older children, trials were generated for infants to test the accuracy of the semi-supervised i-vector based detectors. Figure 3.8 shows the EER values for child, female, and male detectors. In the infants study, acoustic data was only collected from the natural home environment. As was observed in the case of older children dataset, the mean equal error rate across all infants was less than 20%. The mean equal error rate for children was 13.73%, for female speakers it was 18.97%, and for male speakers it was 16.48 %.

3.4 Unsupervised I-vector Based Diarization

In this section, we describe the development of an unsupervised i-vector based diarization system that does not require any labeled data from the recording under investigation. To this end, we will show that it is possible to detect child, adult female, and adult male segments with high accuracy for a completely unlabeled LENA recording.

These methods were developed for the scenario in which no labeled data is present for speakers present in the audio recording. In this scenario, universal older children models, universal infant models, universal female models, and universal male models are used to choose utterances to construct initial subject-specific models. Once these initial subject-specific models are built, semi-supervised recipe is followed.

Algorithm 2 Unsupervised I-vector Training Algorithm

```
Initial Training: 1) Train Universal Child, Female and Male Models.
1. Use universal models to choose child, female, and male segments which have very
high likelihood.
2. Use these chosen segments to train speaker-specific models $S_{models}$ and updated
PLDA.

for fixed number of iterations do
    while parse through unlabeled audio segments do
        Choose audio segments for update using current $S_{models}$ and PLDA
    end while
    Update $S_{models}$ and PLDA
end for
```

3.4.1 Feature Extraction; UBM Training; I-vector Extractor Training

We used the Kaldi toolkit [34] to perform all our experiments. In the initial step, 20 mel frequency cepstrum coefficients (MFCC's) along with their deltas and delta-deltas are computed for every frame of an audio segment at the frame rate of 100/sec. An energy based voice activity detector (VAD) is used to remove frames whose average signal energy is below a threshold. Then, we train a Gaussian Mixture Model of 1024 Gaussians as our

Table 3.9: System Parameter Values Used in the i-vector system

Parameter Name	Value
Acoustic Features	20 MFCC+20 delta+20 delta-deltas
Frame Length	20 ms (Hamming Window)
Frame Rate	100 frames/sec
Universal Background Model	1024 Gaussians
i-vector dimensionality	400

Universal Background Model. Finally, an i-vector extractor is trained such that it maps an audio segment of arbitrary duration into an i-vector of dimension 400.

3.4.2 Universal Child, Adult Female, and Adult Male Detector Models; $PLDA_1$ Training

Universal child, adult female, and adult male detector models were used to select audio segments from an unlabeled audio recording to train subject specific speaker models for that recording. Ideally, one would want to train these models over a significant number of speakers. For our data, we had 20 different patients. So, to detect child, adult female, and adult male segments for an unlabeled audio recording, we used labeled data from other 19 patients to train our universal child, female, and male detector models. The universal models were computed as the mean of i-vectors obtained from audio segments from the 19 patients. For example, suppose $ivec_j^1, ivec_j^2, ..., ivec_j^{N_j}$ are the i-vectors obtained from child audio segments from patient j. Then, the universal child model is given as

$$Universal_{Child} = \frac{\sum_{i=1}^{19}\sum_{k=1}^{N_i} ivec_i^k}{\sum_{i=1}^{19} N_i} \tag{3.16}$$

Similarly, universal adult female and adult male models were computed.

In order to compute a numerical score to determine the likelihood against universal child, adult female and adult male models, $PLDA_1$ scorer was trained. $PLDA_1$ computes the score of two i-vectors $ivec_1$ and $ivec_2$ in the following manner :

$$Score(ivec_1, ivec_2) = \frac{P(ivec_1, ivec_2|H_1)}{P(ivec_1|H_0)P(ivec_2|H_0)} \tag{3.17}$$

where H_1 is the hypothesis that both i-vectors come from child audio segments or adult female audio segments or adult male audio segments or other audio segments and H_0 is the hypothesis that they do not come from the same audio type (i.e, child, adult female, adult male, or other).

3.4.3 Subject Specific Speaker Models; $PLDA_2$ Training

For every unlabeled audio recording, we used the universal child, adult female, and adult male detector models to select audio segments with high likelihood when tested using $PLDA_1$. For example, for computing subject specific child model, we selected segments in the decreasing order of likelihood, starting with the audio segment with the highest likelihood, until the cumulative amount of speech from these selected segments was approximately 1 minute and then computed the mean of these i-vectors. Suppose, audio segments which corresponded to $ivec_1, ivec_2, ..., ivec_N$ i-vectors were selected using universal child model. Then, the subject specific child model was computed as

$$Subject\,Specific_{Child} = \frac{\sum_{i=1}^{N} ivec_i}{N} \tag{3.18}$$

Similar procedure was performed to compute subject specific female and male models. One assumption which holds true for our dataset is that there is not more than one female or male speaking in the audio recording. Another point to note is that even in the presence of another child, the selected 1 minute segments almost always comprised of audio segments belonging to the primary child. Our methods can be extended to the most general case, when there are multiple female or multiple male speakers by introducing a check for the number of clusters present in the selected female or selected male segments, respectively.

$PLDA_2$ was trained using labeled data from 19 patients and audio segments selected from unlabeled audio recording to compute subject specific speaker models. $PLDA_2$ com-

Table 3.10: Average EER Values in Clinic Conditions

Detector Type	EER
Child	14.16
Female	17.63

Table 3.11: Average EER Values in Home Conditions

Detector Type	EER
Child	13.98
Female	15.56
Male	6.78

putes score of two i-vectors $ivec_1$ and $ivec_2$ similar to $PLDA_1$ in the following manner :

$$Score'(ivec_1, ivec_2) = \frac{P(ivec_1, ivec_2|H_1^{')}}{P(ivec_1|H_0')P(ivec_2|H_0')} \tag{3.19}$$

where H_1' is the hypothesis that both i-vectors come from the same speaker and H_0' is the hypothesis that they do not come from the same speaker.

3.4.4 Results for Older Children

In order to test the accuracy of our i-vector based vocalization detectors, we created trials in which the subject specific speaker models were tested against i-vectors obtained from segments not used in training the speaker models and $PLDA_2$ model. The accuracy of the i-vector based vocalization detectors was tested using Equal Error Rate (EER)$\%$. The equal error rate corresponds to the operating point of a detector at which the false rejection rate and false acceptance rate are equal. For clinic conditions, we did not have any male speakers. We see from the tables that the average equal error rates for all detector types in both clinic and home conditions is less than 20$\%$, which suggests accurate child, female and male detection. These results are not as good as the semi-supervised case, where some amount of labeled data is used to train subject specific models, but they are still close enough for actual use.

Table 3.12: Summary of Accuracy of unsupervised i-vector based detectors for infants

Detector Type	Mean EER
Child Home	16.34
Female Home	19.97
Male Home	19.21

3.4.5 Results for Infants

Similar to experiment conducted for older children, trials were generated for infants to test the accuracy of the unsupervised i-vector based detectors. Figure 3.8 shows the EER values for child, female, and male detectors. In the infants study, acoustic data was only collected from the natural home environment. As was observed in the case of older children dataset, the mean equal error rate across all infants were less than 20%. The mean equal error rate for children was 16.34%, for female speakers it was 19.97%, and for male speakers it was 19.21 %

3.5 Summary and Discussion

In this chapter, development of accurate i-vector based diarization methods for LENA audio recordings was done. To this end, two methods to perform accurate diarization were developed. In the first method called as semi-supervised method, 2 minutes of speech was used per speaker to develop initial speaker models and back-end PLDA model. These initial models were then used to annotate unlabeled data. Among unlabeled audio segments, the audio segments which were classified with very high likelihood were chosen to update the initial models. This process was iterated for some small fixed number of iterations to obtain final speaker models and labels. In the second method called as unsupervised method, universal speaker models for child speakers, female speakers, and male speakers were computed. For a completely new unlabeled LENA audio recording, these universal speaker models were used to annotate labels as child, female and male speakers. Among these newly labeled audio segments, a small set of audio segments which were classified into child, female, and male speakers with a very high likelihood were used to compute speaker specific models and then the semi-supervised diarization recipe was followed.

The methods were tested in both clinic conditions where research assistants would interact which children while they were performing a fixed set of tasks, and in home conditions where they were in their natural home environment interacting with their care-givers. In both these conditions, the semi-supervised methods and unsupervised methods were very accurate for all classes of speakers.

CHAPTER 4

APPLICATION OF DIARIZATION METHODS TO COMPUTE LANGUAGE BEHAVIOR AND PRE-LANGUAGE BEHAVIOR STATISTICS

In chapter 3, development of accurate i-vector diarization methods specific to LENA data was described. The main motivation as has been outlined earlier to do accurate diarization is to be come up with language behavior objective measures that can determine if a particular treatment or behavioral intervention is effective for children already diagnosed with autism, and develop pre-language behavior statistics for infants to help diagnose autism early.

In this chapter, a) desirable characteristics of an objective measure, b) computation of utterance rate and its effective in determining autism severity, and d) extracting 5 minute high vocalization window for computing canonical babble statistics for infants is described.

4.1 Need for Objective Measures of Assessment and their Characteristics

One of the main goals in the autism research community is to develop automatic objective measures which can track the efficacy of various treatments. Accurate automatic objective measures, by definition would eliminate subjectivity associated with current methods of assessment based on clinician and caregiver reports. If such measures are computable in home environments, then much more data compared to controlled clinic settings will be available for analysis, enabling more robust estimation of child's language behavior.

Let M denote an objective measure of assessment. Then, any such M should have the following properties:

- *Non-Intrusive (NI):* The sensors used to collect data from an individual diagnosed with ASD to compute M should be as non-intrusive as possible. They should not intrude with individual's normal functioning and should cause minimum inconvenience to the individual.

- *Computable in home environments (H):* M should be computable in naturalistic home conditions. Development of objective measures in naturalistic home conditions would enable analysis of more data for a particular participant than is feasible in controlled clinic settings.

- *Reliable (R):* M should be reliable i.e if the child behavior is similar at any two times t_1 and t_2, then, M measured at t_1 and t_2, denoted by M_{t_1} and M_{t_2} respectively, should be similar.

- *Reveals some aspect of autism severity (AS):* Another important property that any proposed M should have is that it should be able to determine some aspect of autism severity. This could be the overall autism severity score or some specific aspect of social communication or repetitive behavior symptom measures.

Table 4.1: Parameters of semi-supervised i-vector system

Parameter Name	**Parameter Value**
Acoustic features	20 MFCC + 20 delta + 20 delta-deltas
Frame length	20 ms (Hamming Window)
Frame rate	100 frames/sec
UBM	1024 Gaussians
i-vector dimensionality	400
Average child detector EER	9.71

4.2 Utterance Rate Computation and Analysis

4.2.1 Definition

Utterance Rate: Suppose a speaker segmentation algorithm segments an audio recording with multiple speakers into N different segments, separated by speaker boundaries. Let M be the number of segments identified by a child detector to contain child utterances. Then, utterance rate, denoted by U is defined as

$$U = \frac{M}{N}$$

In this study, we used LENA's segmentation algorithm to segment the audio recording based on speaker boundaries. Semi-supervised i-vector based child detector models introduced in [36] were used to accurately detect child segments.

4.2.2 Semi-supervised i-vector based child detectors and utterance rate computation

Figure 1.2 shows the steps involved in computation of utterance rate statistic. As described in the definition of utterance rate, a segmentation algorithm should be used to obtain audio segments. In this work, we used LENA's segmentation algorithm. The human annotators found that there were quite a few segments which had the child and an adult speaking one after the other in the same audio segment. This is an anomaly of the segmentation algorithm. We plan to develop improved segmentation algorithm for our future work. Semi-supervised i-vector based methods were developed to do child and adult voice detection. I-vector based methods have been extensively shown to be the state of the art methods for speaker recognition and diarization tasks in other domains such as telephony, broadcast news, and conversational meetings. Any i-vector based speaker recognition or diarization system has two sub-systems. A front-end which maps an audio utterance of arbitrary duration into a fixed low dimensional vector (usually 100-1000) called as an i-vector which captures both speaker and channel characteristics [23], and a back-end which gives a likelihood score if two i-vectors belong to the same speaker. We used probabilistic linear discriminant analysis (PLDA) scorer as our back-end [24].

Algorithm 3 Semi-Supervised I-vector Training Algorithm

```
Initial Training: 1) Train UBM and I-vector extractor (T) matrix 2) Use labeled information to train initial i-vector speaker models $S_{models}$ and PLDA scorer

for fixed number of iterations do
    while parse through unlabeled audio segments do
        Choose audio segments for update using current $S_{models}$ and PLDA
    end while
    Update $S_{models}$ and PLDA
end for
```

For every subject, we had small amount of labeled data for speakers present in the audio recording. The LENA audio data was single channel data sampled at 16 kHz. We used Kaldi toolkit [34] to run our experiments. A total of 20 MFCC features plus deltas and delta deltas were extracted based on 20 milliseconds (ms) Hamming windowed data with a frame rate of 100/second. A basic energy based voice activity detection was used to select voice only frames. A 1024 Gaussian Mixture Model (GMM) using around 20 hours of data was trained as a Universal Background Model (UBM). An i-vector extractor was trained to generate a 400 dimensional i-vector for a given audio segment. Therefore, every audio segment was mapped into a 400 dimensional i-vector. No labeling information was required to train UBM and i-vector extractor. This is step 1) of initial training part of the algorithm described.

The initial subject specific speaker models for every child and adult were computed using 2 minutes of speech per speaker. To be precise, for any given speaker a set of audio segments was chosen such that the cumulative amount of speech after being processed though VAD in these segments was approximately 2 minutes. The mean of the i-vectors corresponding to these segments was deemed to be the speaker model for that speaker. The data used to train speaker models i.e., 2 minutes of speech per speaker, was used to train the initial PLDA back-end scorer. We used additional data of 2 minutes comprised of toys, TV, environmental noise, silence from every recording. This is step 2) of initial training part of the algorithm described.

We used unlabeled data to retrain our initial speaker and PLDA model updates. The i-vectors corresponding to unlabeled audio segments were tested against initial subject specific trained models trained in step 2) of the initial training using PLDA scorer trained in step 2) of the initial training. Utterances which had very likelihood scores were chosen for updating speaker models and PLDA. The speaker models and the PLDA back-end were then retrained using the initial limited labeled audio segments and the newly chosen utterances for update. This step was repeated for a fixed number of iterations to obtain final subject specific speaker models and PLDA scorer.

In order to test the accuracy of our i-vector based detectors, we created trials in which the speaker models were tested against i-vectors obtained from segments not used in training the speaker models and PLDA scorers. For example, let $Child_1$ denote the speaker model of some subject. Then, $Child_1$ clinic was tested against i-vectors corresponding to child segments, adult segments, toy segments, noise segments, and silence segments from that subject's LENA recording, which were not used in any training. The detection accuracy for child speaker detectors was measured by equal error rate (EER). The mean EER value for all child subjects was 9.71%. In order to compute M in the definition of utterance rate, for every audio segment present in the set of N audio segments, the likelihood of audio segment having a child utterance is computed by testing the i-vector corresponding to

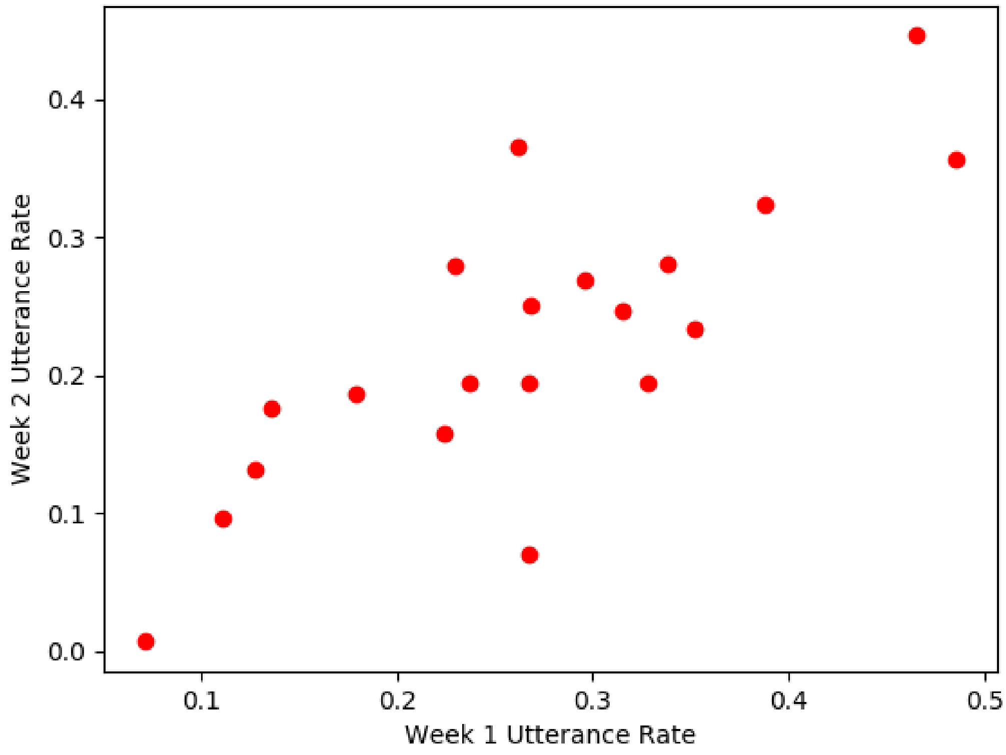


Figure 4.1: Week 1 versus Week 2 Utterance Rate Comparison

the audio segment against the trained i-vector child speaker model. This likelihood score is then compared to a threshold, obtained from testing on small amount of labeled data, and then, a decision is made if the audio segment has a child utterance.

4.2.3 Results and discussion

In sections 4.1 and 4.2, we defined utterance rate and described an i-vector based method to compute it. In this section, we show that utterance rate has the described characteristics NI, H, R, and AS of an objective measure.

The LENA device used to collect audio data is easily accommodated in children's clothing, causing minimal interference to child's normal functioning. Thus, utterance rate computed from this audio data has property NI. In general, for audio-based measures to have

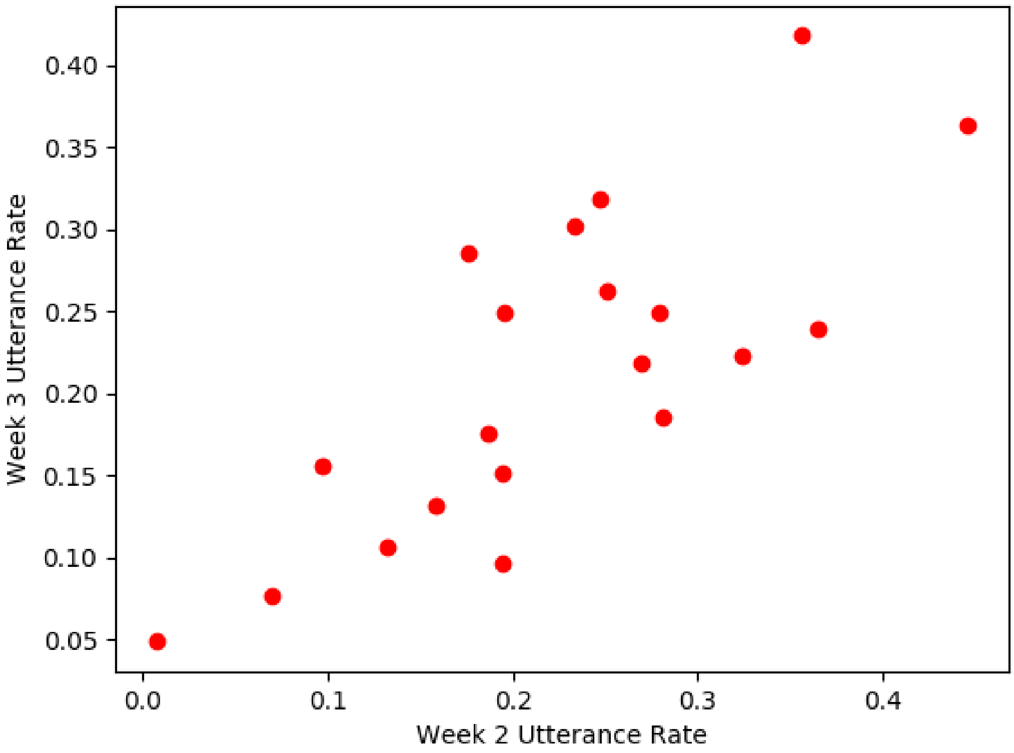


Figure 4.2: Week 2 versus Week 3 Utterance Rate Comparison

the NI property, recording microphones must either be placed near children or be placed on the child clothing without causing much hindrance. For computing utterance rate, no special cues from parents are necessary. As in, it is computed from natural interactions between children and parents. Thus, utterance rate has property H and it can be estimated over more data than measures which can be computed only from special kinds of interactions, such as interactions when parents are instructed to ask specific questions.

As was outlined before, audio data was collected across 3 different weeks, 1^{st}week, 4^{th}week, and 8^{th} week respectively over a period of 8 weeks. For the rest of discussion, we refer to 1^{st} by week 1, 4^{th} week by week 2, and 8^{th} week by week 3 respectively. This is the typical period for which a drug trial lasts. No explicit treatment was administered as part of this study, therefore there is no reason to predict that language behavior should differ

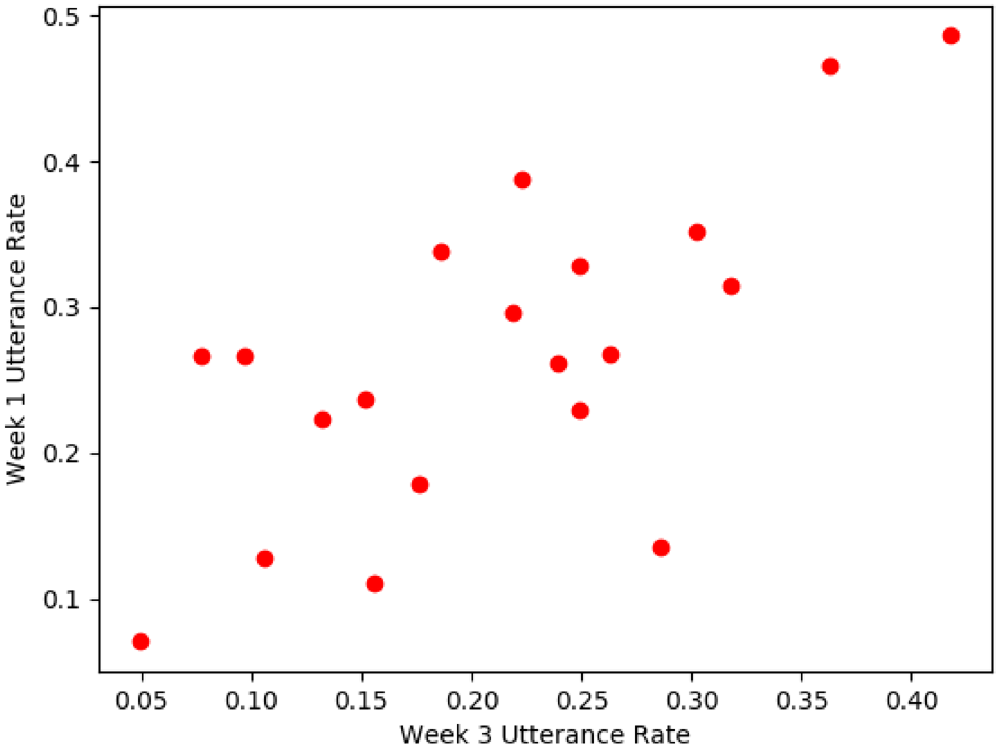


Figure 4.3: Week 3 versus Week 1 Utterance Rate Comparison

across the 8 week period. In order to check if utterance rate has property R, we computed utterance rate for week 1, week 2, and week 3 for all 20 subjects. Figures 4.1, 4.2, and 4.3 show the comparison of utterance rate across different weeks for all 20 subjects. We can infer from figures 4.1, 4.2, and 4.3 that utterance rate is relatively consistent across three different instances of measurement. The Pearson correlation coefficient (ρ) between week 1 and week 2 was 0.79, between week 2 and week 3 was 0.77, and between week 3 and week 1 was 0.70 respectively. Thus, utterance rate has property R.

The subjects in this investigation as noted earlier were diagnosed with autism spectrum disorder ASD. The amount of autism severity, however, varied from one participant to another. The amount of autism severity was measured by ADOS-CSS (autism diagnostic observation schedule calibrated severity score) as well as CSS-SA (calibrated severity

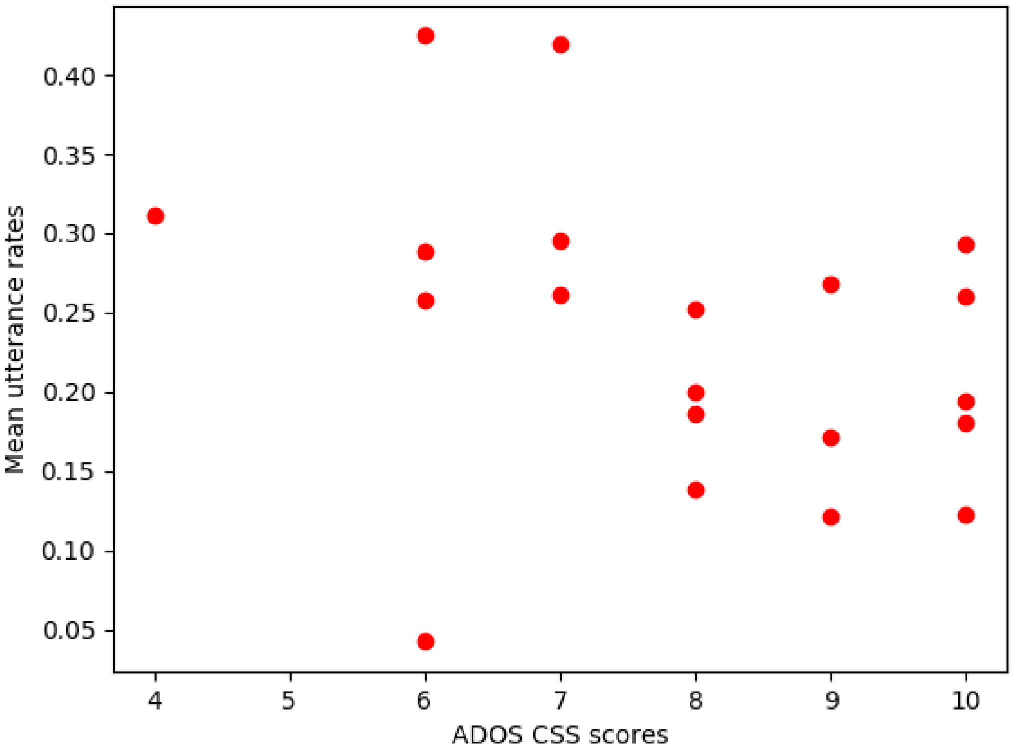


Figure 4.4: Comparison of utterance rate with ADOS CSS scores

score social affect) score which is specifically targeted towards atypical communication symptoms of autism [37]. These scores are discrete-valued from $0 - 10$ with 10 being the highest severity level and are computed from raw ADOS scores after examination by a licensed clinician. These measures are derived from clinical judgment. These measures capture many different aspects of social communication deficits in ASD including gaze, and gestures in addition to language challenges. There were 6 different ADOS-CSS and CSS-SA (4,6,7,8,9,10) values in our sample. Figures 4.4 and 4.5 show the comparison of mean utterance rate computed with ADOS CSS scores and CSS SA scores respectively. Figure 4.4 shows that utterance rate can possibly classify subjects into two classes: Class a) subjects with high ADOS CSS score (8,9,10) and Class b) subjects with low ADOS score (4,6,7). However, within a particular class, let's say of high ADOS CSS scores, it

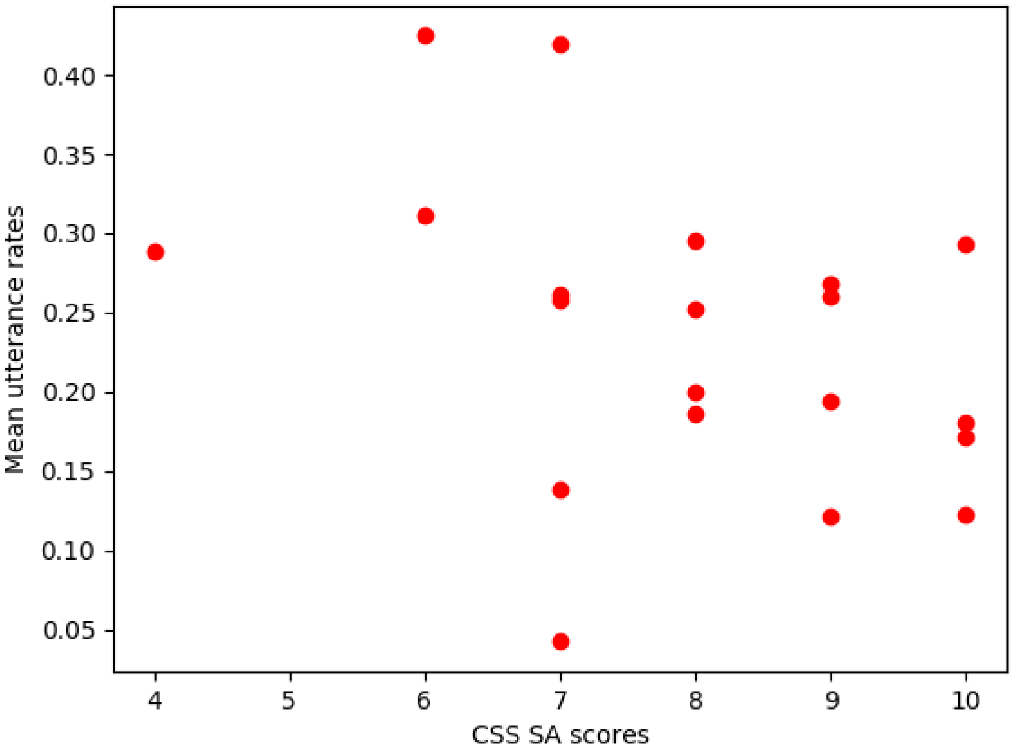


Figure 4.5: Comparison of utterance rate with CSS SA scores

does not have the capability to distinguish between 8,9, and 10 ADOS CSS scores. The average utterance rate for subjects with low ADOS CSS score was 0.2877 while for subjects with high ADOS CSS score was 0.199, suggesting it can be used to determine if a subject belongs to the higher autism severity class or lower autism severity class. Thus, utterance rate can reveal some information about autism severity and has property AS.

One possible improvement is development of language based measures which factor in child-parent interaction. These might be more descriptive in revealing autism severity. Also, more measurements of utterance rate or any proposed objective measure per subject will make it possible to study the distribution of that statistic and study its dynamics over time if any treatment is administered. Another improvement could be increased number of subjects. This will enable a more rigorous analysis of a proposed objective measure.

4.3 Extraction of 5 Minute High Vocalization Child Window for Canonical Babble Estimation

One of the main goals of autism research as has been highlighted before, is to diagnose autism spectrum disorder at an early age. The present clinical diagnostic tools for autism spectrum disorder can diagnose autism earliest at the age of 24 months. The current consensus among autism researchers is that if diagnosis of ASD could be done at an earlier stage, then, treatments and behavioral intervention could be more effective.

In order to enable early diagnosis of ASD, it is therefore important to be able to characterize pre-language development and paralinguistic behavior of children before 24 months. Some possible measures of characterizing pre-language child development and paralinguistic behavior are canonical babble ratio; paralinguistic behavior such as crying, laughing, screaming; and motherese. In this section, a method is described to compute canonical babble ratio using human intervention for day-long naturalistic audio recordings.

4.3.1 Canonical Babble Ratio

Canonical Babbling (CB) is an important speech-language milestone in the first year of a child's life. It comprises of canonical syllables, defined as fully articulated sound sequences with a consonant-like and vowel-like sound, with a rapid transition between them [38], eg. ba-ba-ba, ga-ga-da-ba etc. It is observed in typically developing infants between 7-10 months [39–41]. Analysis of canonical syllables produced by infants is of interest because canonical syllable production has been shown to relate to later speech-language abilities [42, 43]. It is conjectured that CB may serve as an early developmental marker in differentiating neurodevelopmental disorders [44, 45]. A canonical syllable consists of all of the following elements:

1. a mature and clear vowel-like sound.
2. a mature and clear super-glottal consonant-like sound.

3. coupling of 1 and 2 in a time manner, i.e rapid transition between consonant-like sound and vowel-like sound.

4. high speech-like quality.

Non-canonical syllables (Non-CS) consist of only vowel-like sounds or vowel-like sounds with unclear consonant-like sounds or isolated consonants (eg. mmmmmm). Vegetative sounds like burping, sneezing, coughing; raspberries or trills eg. (brrr); cooing, gurgling, comfort sounds; crying, whining, screaming; squealing (vocalizations at high pitch); growling; and breathing sounds are not considered child vocalizations.

Canonical Babble Ratio (CBR): Suppose, an audio window consisting of N_1 canonical syllables and N_2 non-canonical syllables is given. Then, the canonical babble ratio (CBR) is given by

$$CBR = \frac{N_1}{N_1 + N_2} \tag{4.1}$$

Here, $N_1 + N_2$ are the total number of child vocalizations. For canonical babble ratio to be a reliable measure, it should stabilize after some threshold number of child vocalizations. That is, if $x \in \mathbb{N}$ is number of child vocalizations, then, $CBR(x)$ should more or less be constant, for $x > M$, where M is the threshold number of child vocalizations.

Previous studies which analyzed short video clips used 5 minute of audio window to compute the canonical babble ratio (CBR) [46, 47]. Currently, there are no automatic accurate methods to determine if a given child vocalization is a canonical syllable or not. So, human annotation is the only way to analyze them.

LENA audio recordings in the IBIS dataset are 16 hour day long recordings. The distribution of child vocalizations is not uniform across the 16 hour period. For example, there are large periods of time in the 16 hour recording when no child vocalizations are recorded (eg. the child is sleeping). An important problem therefore is to extract an audio window from the 16 hour recording which has significant number of child vocalizations. Since, the previous studies [46, 47] used 5 minutes of audio window to compute CBR, the problem is: given a 16 hour LENA recording, extract a 5 minute audio window which has high number of child vocalizations.

Algorithm 4 Algorithm to extract 5 minute high child vocalization window from 16 hour day long recording

```
Given: 16 hours audio recording $AR$ and a diarization scheme $D$
1. Run $D$ on $AR$.
2. Divide $AR$ into contiguous regions of approximately 5 seconds. Maintain speaker boundaries from D when dividing $AR$ into contiguous regions.
3. Compute number of child segments per contiguous region.
4. Sort the contiguous regions according to number of child segments present in descending order.
5. Select the top 60 contiguous regions.
```

Ideally, i-vector based diarization scheme should have been used in Algorithm 3. to extract the 5 minute high child vocalization audio window. However, the CBR experiments were performed before the development of i-vector based diarization scheme for IBIS data. Therefore, LENA diarization output was used to illustrate how a diarization scheme can be used to obtain 5 minute high child vocalization audio window from 16 hour day long recordings using Algorithm 3. Since, LENA diarization is not accurate as shown in chapter 2, there most likely exist other 5 minute audio windows during the 16 hour recordings which have a higher number of child vocalizations. But, since the 5 minute audio window is going to be hand annotated for computing canonical syllables and non-canonical syllables, as long as there are enough child vocalizations, a reliable CBR can be computed.

Table 4.2: Canonical Babble Ratio

	HR-ASD	HR-Neg	LR
	N=11 (8 males)	N=35 (22 males)	N=26 (15 males)
	Mean (SD)	Mean (SD)	Mean (SD)
Canonical Babble Ratio	0.49 (0.32)	0.53 (0.38)	0.67 (0.45)
Canonical Babble Syllables	62.73 (50.44)	58.51 (40.26)	74.15 (53.98)
Child Vocalization Counts	120.45 (47.92)	114.51 (29.68)	107.88 (24.76)

For a total of 72 subjects in the IBIS dataset, canonical babble ratio was computed. Among the 72 children, 11 were high-risk-ASD (HR-ASD), 35 were high-risk-negative

(HR-Neg), and 26 were low risk. The 5 minute audio window was provided to human annotators using a modified version of the GUI annotation toolkit described in chapter 2. For every audio segment, annotators had to annotate if a child speech like vocalizations, child non-speech sounds, or other child sounds were present. If child speech like vocalizations were present, then, they were further labeled as canonical syllables or non-canonical syllables. The number of child vocalizations in both the cases was recorded. Additionally, the exact child vocalization was transcribed in case of canonical syllables.

Table 4.2 shows the computed values of canonical babble ratios for the 72 subjects. The mean CBR for the HR-ASD group was the lowest (0.49), while mean CBR for the LR was the highest (0.67). However, the group differences are not significant. This suggests that a more deeper analysis of canonical babble ratio is required to deduce whether canonical babble ratio can act as an early marker for diagnosing ASD. Some possible lines of investigation could be to track the trajectory of canonical syllable production longitudinally over time. These trajectories might reveal more insight rather than an estimate of canonical babble ratio at a specific time in the early-language development stage.

4.4 Summary and Discussion

In this chapter, the need for objective measurements of assessment for children diagnosed with ASD was addressed. A formal list of properties that an objective measure of assessment should have was introduced. An objective measure of assessment called utterance rate was developed using the semi-supervised i-vector child detectors described in chapter 3. It's reliability was analyzed across 3 weeks and it was shown that it has potential to differentiate between subjects with high ADOS CSS score (8,9,10) and subjects with low ADOS CSS score (4,6,7). Finally, a method to extract a high child vocalization audio window of 5 minutes from 16 hr day long recording was developed and its application to computing canonical babble statistics was studied.

CHAPTER 5
CONCLUSIONS

5.1 Thesis Contributions

1. LENA device has become one of the most widely used devices to collect audio recordings for autism research studies. The LENA device comes with a software which does automatic computation of many language behavior statistics based on diarization. Many autism research studies use these statistics at face-value without critically evaluating whether these statistics are accurate enough for the dataset which is being studied in the research study. In this work, LENA algorithms were evaluated and it was shown that they are not accurate for older children (aged 5-17 years) as well infants (aged 9 months and 15 months).

2. Two accurate i-vector based diarization algorithms were developed for LENA data. One called the semi-supervised i-vector based diarization for the scenario in which small amount of labeled data was available per speaker, and second, called the unsupervised i-vector based diarization for the scenario in which no label data was available for the LENA audio recording to be diarized.

3. Semi-supervised i-vector diarization was applied to compute a possible objective measurement called the utterance rate. It was shown to be reliable across 3 different weeks and could differentiate between subjects with high ADOS CSS score (8,9,10) and subjects with low ADOS score CSS (4,6,7). A method to extract 5 minutes of high child vocalization audio window from a 16 hour day recording was shown and canonical babble ratio computed.

5.2 Possible Future Research Directions

There are many possible directions in which the work can be extended. Some of them are listed below:

1. Develop a more accurate segmentation algorithm than LENA's segmentation. This would require labeling of segment boundaries at the order of frame wise annotation, to train and test algorithms.

2. Use of deep neural network based generative models to generate fixed dimensional speaker embeddings instead of i-vector factor analysis framework to generate low fixed-dimensional representation of audio utterances.

3. Development of automatic speech recognition (ASR) methods for LENA data. ASR is a well studied problem in other domains and therefore, availability of accurate transcribed LENA data should lead to development of ASR methods. A less well studied problem is to automatically characterize development of pre-language for infants over time. This would lead to automatic estimation of canonical and non-canonical syllables.

4. Development of automatic methods for motherese or child-directed speech for infants.

REFERENCES

[1] Jon Baio et al. “Prevalence of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, United States, 2014”. In: *MMWR Surveillance Summaries* 67.6 (2018), p. 1.

[2] Michael Ford et al. “The LENA language environment analysis system: Audio specifications of the DLP-0121”. In: *Retrieved online* (2008).

[3] Charles R Greenwood et al. “Automated Language Environment Analysis: A Research Synthesis”. In: *American journal of speech-language pathology* 27.2 (2018), pp. 853–867.

[4] Sheila Degotardi, Feifei Han, and Jane Torr. “Infants experience with near and cleareducator talk: individual variation and its relationship to indicators of quality”. In: *International Journal of Early Years Education* (2018), pp. 1–17.

[5] Rachel R Romeo et al. “Beyond the 30-Million-Word Gap: Childrens Conversational Exposure Is Associated With Language-Related Brain Function”. In: *Psychological science* 29.5 (2018), pp. 700–710.

[6] Paul J Yoder et al. “Stability and validity of an automated measure of vocal development from day-long samples in children with and without autism spectrum disorder”. In: *Autism Research* 6.2 (2013), pp. 103–107.

[7] Dwight W Irvin et al. “Child and classroom characteristics associated with the adult language provided to preschoolers with autism spectrum disorder”. In: *Research in Autism Spectrum Disorders* 7.8 (2013), pp. 947–955.

[8] Sloane Burgess, Lisa Audet, and Sanna Harjusola-Webb. “Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with ASD”. In: *Journal of Communication Disorders* 46.5-6 (2013), pp. 428–439.

[9] Anne S Warlaumont et al. “A social feedback loop for speech development and its reduction in autism”. In: *Psychological science* 25.7 (2014), pp. 1314–1324.

[10] Tiffany Woynaroski et al. “The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder”. In: *Autism Research* 10.3 (2017), pp. 508–519.

[11] Adrian Garcia-Sierra, Nairan Ramírez-Esparza, and Patricia K Kuhl. "Relationships between quantity of language input and brain responses in bilingual and monolingual infants". In: *International Journal of Psychophysiology* 110 (2016), pp. 1–17.

[12] Virginia A Marchman et al. "Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure". In: *Developmental science* 20.1 (2017), e12425.

[13] Eon-Suk Ko et al. "Entrainment of prosody in the interaction of mothers with their young children–ERRATUM". In: *Journal of child language* 43.4 (2016), pp. 964–965.

[14] Anna V Sosa. "Association of the type of toy used during play with the quantity and quality of parent-infant communication". In: *JAMA pediatrics* 170.2 (2016), pp. 132–137.

[15] Mark VanDam et al. "HomeBank: An online repository of daylong child-centered audio recordings". In: *Seminars in speech and language*. Vol. 37. 2. NIH Public Access. 2016, p. 128.

[16] Rebecca M Jones et al. "Placebo-like response in absence of treatment in children with Autism". In: *Autism Research* 10.9 (2017), pp. 1567–1572.

[17] Rebecca Grzadzinski et al. "Parent-reported and clinician-observed autism spectrum disorder (ASD) symptoms in children with attention deficit/hyperactivity disorder (ADHD): implications for practice under DSM-5". In: *Molecular autism* 7.1 (2016), p. 7.

[18] Fredric J Harris. "On the use of windows for harmonic analysis with the discrete Fourier transform". In: *Proceedings of the IEEE* 66.1 (1978), pp. 51–83.

[19] Douglas A Reynolds. "A Gaussian mixture modeling approach to text-independent speaker identification." In: (1993).

[20] Douglas A Reynolds, Richard C Rose, et al. "Robust text-independent speaker identification using Gaussian mixture speaker models". In: *IEEE transactions on speech and audio processing* 3.1 (1995), pp. 72–83.

[21] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. "Speaker verification using adapted Gaussian mixture models". In: *Digital signal processing* 10.1-3 (2000), pp. 19–41.

[22] Leonard E Baum et al. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* 41.1 (1970), pp. 164–171.

[23] Najim Dehak et al. “Front-end factor analysis for speaker verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

[24] Simon JD Prince and James H Elder. “Probabilistic linear discriminant analysis for inferences about identity”. In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE. 2007, pp. 1–8.

[25] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson. “Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, pp. 4257–4260.

[26] Patrick Kenny. “Bayesian speaker verification with heavy-tailed priors.” In: *Odyssey*. 2010, p. 14.

[27] Seyed Omid Sadjadi et al. “The 2016 NIST Speaker Recognition Evaluation.” In: *Interspeech*. 2017, pp. 1353–1357.

[28] *NIST 2018 Speaker Recognition Evaluation Plan*. 2018. (Visited on 08/04/2019).

[29] *NIST 2019 Speaker Recognition Evaluation Plan*. 2019. (Visited on 08/04/2019).

[30] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel. “Eigenvoice modeling with sparse training data”. In: *IEEE transactions on speech and audio processing* 13.3 (2005), pp. 345–354.

[31] Roland Kuhn et al. “Eigenvoices for speaker adaptation”. In: *Fifth International Conference on Spoken Language Processing*. 1998.

[32] Patrick Kenny et al. “Joint factor analysis versus eigenchannels in speaker recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1435–1447.

[33] Patrick Kenny et al. “A study of interspeaker variability in speaker verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008), pp. 980–988.

[34] Daniel Povey et al. “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. EPFL-CONF-192584. IEEE Signal Processing Society. 2011.

[35] Maryam Najafian et al. “Employing speech and location information for automatic assessment of child language environments”. In: *SPLINE*. 2016.

[36] Rahul Shivaji Pawar et al. “I-vector based diarization for LENA data and its application in computing language behavior statistics for individuals with autism”. In: *Journal of Acoustical Society of America* (Submitted).

[37] Katherine Gotham, Andrew Pickles, and Catherine Lord. “Standardizing ADOS scores for a measure of severity in autism spectrum disorders”. In: *Journal of autism and developmental disorders* 39.5 (2009), pp. 693–705.

[38] D Kimbrough Oller. *The emergence of the speech capacity*. Psychology Press, 2000.

[39] Rebecca E Eilers and D Kimbrough Oller. “Infant vocalizations and the early diagnosis of severe hearing impairment”. In: *The Journal of pediatrics* 124.2 (1994), pp. 199–203.

[40] Liselotte Roug, Ingrid Landberg, and L-J Lundberg. “Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life”. In: *Journal of child language* 16.1 (1989), pp. 19–40.

[41] Eugene H Buder, Anne S Warlaumont, and D Kimbrough Oller. “An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective”. In: *Comprehensive perspectives on child speech development and disorders: Pathways from linguistic theory to clinical practice* 4 (2013), pp. 103–134.

[42] John L Locke and Dawn M Pearson. “Linguistic significance of babbling: Evidence from a tracheostomized infant”. In: *Journal of Child Language* 17.1 (1990), pp. 1–16.

[43] Ken M Bleile, Rachel E Stark, and Joy Silverman McGowan. “Speech development in a child after decannulation: Further evidence that babbling facilitates later speech development”. In: *Clinical Linguistics & Phonetics* 7.4 (1993), pp. 319–337.

[44] Katie Belardi et al. “A retrospective video analysis of canonical babbling and volubility in infants with Fragile X syndrome at 9–12 months of age”. In: *Journal of autism and developmental disorders* 47.4 (2017), pp. 1193–1206.

[45] D Kimbrough Oller et al. “Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development”. In: *Proceedings of the National Academy of Sciences* 107.30 (2010), pp. 13354–13359.

[46] Elena Patten et al. “Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency”. In: *Journal of autism and developmental disorders* 44.10 (2014), pp. 2413–2428.

[47] Rhea Paul et al. “Out of the mouths of babes: Vocal production in infant siblings of children with ASD”. In: *Journal of Child Psychology and Psychiatry* 52.5 (2011), pp. 588–598.