Workshop on Child Computer Interaction
6-7 Sep 2016, San Francisco, USA

Automatic measurement and analysis of the child verbal communication using classroom acoustics within a child care center

Maryam Najafian[1], *Dwight Irvin*[2], *Ying Luo*[2], *Beth S. Rous*[2], *John H. L. Hansen*[1]

[1]Center for Robust Speech Systems
University of Texas at Dallas, Richardson, TX, USA
[m.najafian,john.hansen]@utdallas.edu

[2]College of Education,
University of Kentucky, Lexington, KY, USA
[dwight.irvin,ying.luo,beth.rous]@uky.edu

Abstract

Understanding the language environment of early learners is a challenging task for both human and machine, and it is critical in facilitating effective language development among young children. This papers presents a new application for the existing diarization systems and investigates the language environment of young children using a turn taking strategy employing an i-vector based baseline that captures adult-to-child or child-to-child conversational turns across different classrooms in a child care center. Detecting speaker turns is necessary before more in depth subsequent analysis of audio such as word count, speech recognition, and keyword spotting which can contribute to the design of future learning spaces specifically designed for typically developing children, or those at-risk with communication limitations. Experimental results using naturalistic child-teacher classroom settings indicate the proposed rapid child-adult speech turn taking scheme is highly effective under noisy classroom conditions and results in 27.3% relative error rate reduction compared to the baseline results produced by the LIUM diarization toolkit.

Index Terms: child speech, speech turn taking, language environment analysis

1. Introduction

The quality and number of interactions that accompany a rich language environment contribute to essential language developmental outcomes in early childhood [1]. For humans, analyzing the large quantity of data is not practical and building real-time solutions that provide actionable analysis is cost-prohibitive. On the other hand for machines, scaling to process large quantities of data is possible but there is a need to develop robust speech processing and location tracing systems that can bring consistency and reliability to the analysis. In this study, we employ the LENA recording device [2, 3] and robust analytical algorithms to lay the foundation for a machine-based solution.

In this study we recorded and tracked the location of 33 children of age 2.5 to 5 years old across 4 classrooms in a high-quality childcare center in the United States at various time points during the day. We aim to determine how much of the child's interaction involves other children and how much is from classroom teachers. For this purpose a speech activity detector followed by a diarization is required to be able to detect fast turn changes during child-adult conversations.

Our motivation is to automate assessment of child's language environment, which may assist automatic monitoring of child language acquisition and development progress. In this study we describe the current state of the algorithms applied to similar tasks and their drawbacks for our current application. next, we propose a system which addresses challenges and categorizes every 1.5 seconds of data into four main categories: (i) speech produced by the child (ii) speech directed towards the child by an adult, (iii) speech directed towards the child by another child and (iv) non-speech (the stream of background music or conversation by other adults and children). Next we provide an example analysis carried out on the data from a typical children classroom (e.g 15-20 students and 1-3 teachers) at three different time points. Finally, we present the confusion matrix for the child-adult turn taking system and present a conclusion and layout of our future work.

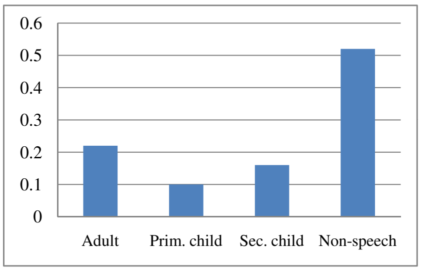


Figure 1: *Percentage of different classes of data in the database*

2. Speech database description

For speech data collection a light weight compact digital audio recorder (LENA device [2, 3]) is worn by 33 children of age 2.5 to 5 years old. The audio is recorded throughout a typical day at a childcare center, during at least one of the three different time points, where the subject was participating in different activities. We used 4.5 hours of audio recording gathered by the LENA unit attached to 18 children (approximately 15 minutes each) to train our speech analysis systems. In our experiments we used a 3-fold cross validation so no speaker appeared simultaneously in the training and test sets.

Table 1: *Ground-truth analysis for the dataset*

Segment class	Average duration	Average turn duration
Primary child	1.9s	1.8s
Secondary child	1.8s	1.6s
Adult	2.2s	2.1s

For system evaluation purpose, this data was partitioned into approximately 1.5 second segments and each cut was labeled correspondingly. From the manual labels gathered, we es-

http://dx.doi.org/10.21437/WOCCI.2016-10

timated that 52%, 22%, 10%, and 16% of our speech database belongs to non-speech, adult speech, secondary child speech, and primary child speech categories respectively (Figure 1).

1. **Non-speech**: the stream of background noise, silence, music or conversation produced by other children or adults more than 8 feet away from the primary child
2. **Primary child**: speech initiated by the child wearing the LENA unit
3. **Secondary child**: speech originated by other children and directed at the primary child within his/her close proximity
4. **Adult**: speech originated by an adult and directed at the primary child within his/her close proximity

Table 1 reports average segment and turn durations within the database. The average segment duration for each label refers to the average time during which a certain class is active. Conversely, the average turn duration refers to the average time during which there is no change in segment activity and is thus always smaller than the average speaker duration.

3. Related work

Speaker diarization is the task of identifying "who spoke when" in an audio stream containing multiple speakers. Studies on state-of-the-art diarization systems have isolated three main issues: overlapping speech, effects of background noise, and speech/non-speech detection errors on clustering with significant performance variance across systems and data stream types [4, 5, 6]. The state-of-the-art system for broadcast news speaker diarization is composed of 5 steps. First, music and jingle regions are removed using Viterbi decoding. Next, an acoustic segmentation followed by a Hierarchical Agglomerative Clustering (HAC) splits and then groups the signal into homogeneous parts according to speakers and background. In this step, each segment or cluster is modeled by a Gaussian distribution with a full covariance matrix, and the Bayesian Information Criterion (BIC) [7] is employed both as similarity measure and as stop criterion. Then, a Gaussian Mixture Model (GMM) is trained for each cluster via the Expectation-Maximization (EM) algorithm. The signal is then re-segmented through a Viterbi decoding. The system finally performs another HAC, using the Cross-Likelihood Ratio (CLR) [8] measure and GMMs trained with the Maximum A Posteriori algorithm (MAP) [9]. Using this diarization routine, several broadcast news and meeting diarization toolkits have proposed in the literature, namely the CMU Segmentation tool [10], the LIUM open-source speaker diarization toolbox [11], the AudioSeg Audio segmentation toolkit [12], the speaker diarization and recognition library ALIZE [13], the SHoUT diarization toolkit [14], the diarization system by LIA and CLIPS laboratories [15] in which segmentation and clustering are done iteratively and jointly. Later, IDIAP published DiarTK [16] where clustering and segmentation are based on the information bottleneck principle. After the success of i-vector features in different fields of speaker verification, IDIAP applied Integer Linear Programming (ILP) based clustering to the i-vector features [17]. Recently, Yella [4] proposed an approach based on Information Bottleneck with Side Information (IBSI) based diarization to suppress artifacts of background noise and non-speech segments introduced during spontaneous communication clustering. These systems can perform better when there are pauses between conversational turn takings rather than spontaneous speech. Our child-adult turn taking application is very similar to the speech diarization systems, as it requires to find turning points as the source of the audio segment changes [18].

4. System description

In this section we start with describing an i-vector based child-adult diarization system with a Support Vector Machine (SVM) [19] clasifier and a Threshold Optimized Speech Activity Detector (TO-COMBO-SAD [20]). Then we describe a standard speaker diarization (LIUM speaker diarization toolkit [11]).

4.1. Proposed i-vector based child-adult speech turn taking system

In this section we describe our threshold optimized i-vector based language environment classification system with 28% classification error rate in categorizing the audio (with intense time-varying acoustic noise) into audio segments under four categories of non-speech, primary child, secondary child, and adult speech. To assist in removing durations of non-speech that are common in naturalistic audio recordings and focus the diarization on speech regions alone, we exploited a Threshold Optimized SAD system (TO-Combo-SAD) [20]. This unsupervised SAD technique is designed to be noise robust and has been particularly effective in multiple DARPA RATS evaluations [21, 22].

The idea behind the text-independent (unsupervised) i-vector based approach is initiated from the Joint Factor Analysis (JFA) [23] technique proposed for speaker verification. In JFA, speaker and channel variabilities are represented in two separate subspaces, whereas in the i-vector approach only one single space is defined for all types of variability (including both speaker and session variabilities). This is called the 'total variability' space. The motivation for the use of a single 'total variability', space is that in the JFA traces there are speaker dependent information, found in the channel factors, and therefore separating speaker and channel variabilities will lead to a loss of some speaker dependent information [24].

I-vectors provide a low-dimensional representation of feature vectors that can be successfully used for classification and recognition tasks. I-vectors were initially introduced for speaker recognition for adults [25, 24] and children [26, 27]. After their success in this area, they were applied to language recognition [28], and accent recognition [29, 30, 31, 32] areas. The proposed diarization system was inspired by the success of i-vectors in age-group identification task for children and adults [33] and child-adult speaker diarization system proposed by Najafian et al. [34] for longer duration segments (3 seconds). Our child-adult turn taking strategy for audio segments of 1.5 seconds is summarized in Figure 2.

The architecture of an i-vector system with a Support Vector Machine (SVM) [19] classifier and a TO-Combo-SAD unit is illustrated in Figure 3. The process of building an i-vector system consists of the following stages.

TO-Combo-SAD: Combo-SAD computes five noise robust features at the frame level for each audio segment and projects the combined feature vectors into a single 1-dim space using Principal Component Analysis. Let f_i be the Combo feature vector for the i^{th} frame and $\bar{f}_i$ is the normalized feature vector (mean 0 and variance 1). Now, let X be the principal eigenvector(corresponding to the largest eigenvalue of the feature covariance matrix). Finally, let p_i be the projection of $\bar{f}_i$ on X. The Combo features are designed to have higher values for speech and lower values for noise/background. Therefore, p_i value will be generally higher for speech than background. The Combo-

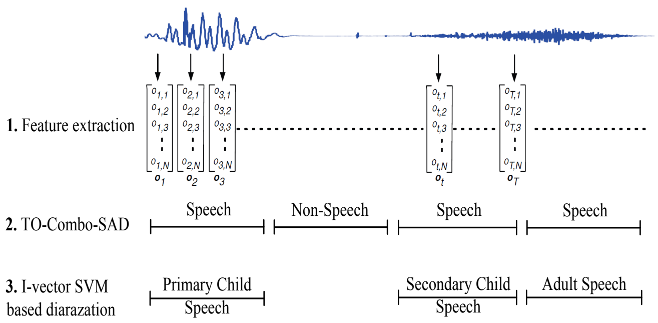

1. Feature extraction
$o_{1,1}$ $o_{1,2}$ $o_{1,3}$ $o_{1,N}$ o_1
$o_{2,1}$ $o_{2,2}$ $o_{2,3}$ $o_{2,N}$ o_2
$o_{3,1}$ $o_{3,2}$ $o_{3,3}$ $o_{3,N}$ o_3
$o_{t,1}$ $o_{t,2}$ $o_{t,3}$ $o_{t,N}$ o_t
$o_{T,1}$ $o_{T,2}$ $o_{T,3}$ $o_{T,N}$ o_T
2. TO-Combo-SAD
Speech
Non-Speech
Speech
Speech
3. I-vector SVM based diarazation
Primary Child Speech
Secondary Child Speech
Adult Speech

Figure 2: *Main stages of the child-adult turn taking system*

SAD exploits this principle for decision making. It trains a two-mixture GMM (Gaussian Mixture Model) on-the-fly to automatically determine the speech and background clusters. The mixture with larger mean value is hypothesized to belong to speech and vice-versa. In order to build an effective speech model, we first train a large mixture GMM on speech data extracted from annotated corpora (sources are Switchboard, and Fisher). Next, the means of this GMM are projected into the Combo SAD's single-dimension decision making space, where m_j is the j^{th} mixture mean of the M-mixture GMM, and $\hat{m}_j$ is the corresponding projected value. Let μ_{hs} and μ_{hp} be the hypothesized speech and background mixture means of the GMM, and let μ_{ts} be the mean of projected values $\hat{m}_j$. Here $\hat{m}_j$ represents the prior model of speech (since it was built with speech data from annotated corpora), while μ_{ts} can be viewed as the posterior model of speech (since it is built by Combo-SAD from data). If $\mu_{hs} \geq \mu_{ts}$, then we trust the posterior model of speech and use it for decision making. Alternatively, if $\mu_{hs} < \mu_{ts}$, then we use the prior model of speech for decision making.

In the next step, the mixture means are used to compute the SAD threshold and speech/pause decisions are made. The threshold value τ is computed using a simple convex combination, where w is the weight factor such that $0 \leq w \leq 1$.

$$\tau = w \max(\mu_{hs}, \mu_{ts}) + (1 - w)\mu_{hp} \qquad (1)$$

The proposed method of estimating the threshold value has a significant impact on performance of SAD on speech-sparse and non-speech regions. The labels for non-speech segments are recorded at this stage and only the segments that are marked as speech are given as inputs to the i-vector system for both training and test.

Feature extraction: The speech is pre-emphasised with periods of non-speech discarded using TO-Combo-SAD. Speech is then segmented into 25-ms frames with a shift of 10-ms between frames, and a Hamming window applied to each frame. The short-time magnitude spectrum, obtained by applying the FFT, is passed to a bank of 27 Mel-spaced triangular band-pass filters. Each speech frame is then represented as a 42-dimensional Mel Frequency Cepstral Coefficients (MFCCs) feature vectors consisting of 0$^{\text{th}}$- to 12$^{\text{th}}$-order Cepstral coefficients, log energy, and all delta and delta-delta variants.

UBM construction: Speech from the training subset is used to estimate the parameters of a the Universal Background Model (UBM).

Baum-Welch statistics: The UBM trained in the previous stage can now be used for extracting the zero- and first-order Baum-Welch statistics centralised over the UBM mean.

Total variability modeling: Each utterance is described in terms of a speaker and channel dependent GMM mean supervector M, where $M = m + Tw$. Suppose that K is the number of Gaussian components in the UBM, and F is the dimension of the acoustic feature vectors. The speaker and channel independent statistics, m, of dimension $KF \times 1$ is constructed by concatenating the means for each Gaussian component of the UBM. The aim of the total variability modelling technique is to find a low rank rectangular 'total variability matrix' (T-matrix), T, of dimension $KF \times H$ with $H \ll KF$, and low dimensional 'identity vector', w, of dimension $H \times 1$ such that the probability of the training utterances given the model defined by the supervector M is maximized. Given each utterance from the corresponding child or adult speaker, the value of T-matrix and i-vector are estimated iteratively using EM to maximize the likelihood of the training data. In the Expectation step, T is assumed to be known, and we update w. In the Maximization step, w is assumed to be known and we update T.

Extracting the i-vectors: For the utterance dependent mean offset, Tw, the components of the i-vector best describe the coordinates of the utterance in the reduced total variability space. Given the utterance, u, in the Expectation step, the i-vector w which is the mean of posterior distribution is updated using the current value of the T-matrix, and the Baum-Welch statistics extracted from the UBM. Presenting the utterances in the low-dimensional total variability space, ensures that for representing a new speaker only a small number of parameters need to be estimated. To achieve this the total variability space needs to encapsulate as much as possible of the supervectors in its restricted number of dimensions.

SVM: Given a set of labeled training utterances from 3 classes, our multi-class SVM classifier is a collection of 2-class SVMs with linear Kernel, which is trained using the corresponding accent-specific i-vectors. Next, the test speaker's i-vector is scored against each SVM (using a 'one against all' approach). The class which gives the maximum score determines the label of the test utterance. In our system, the UBM was trained on the training subset of the dataset using various number of UBM components and T-matrix ranks (3-fold croos validation). Our system uses a UBM with 256 components, and a T-matrix of rank 25 (chosen empirically).

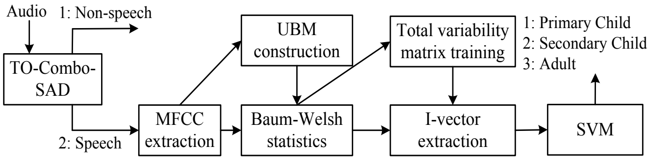


Figure 3: *Proposed diarization system*

4.2. LIUM speaker diarization toolkit

In this section we describe the application of the LIUM speaker diarization toolkit [11] for child-adult and non-speech classification with 38.5% error rate. This diarization system is composed of acoustic BIC segmentation followed by BIC [10] hierarchical agglomerative clustering. The Universal Background Model (UBM) is adapted (Maximum A Posteriori MAP) for each cluster. A cluster is modeled by a Hidden Markov Model (HMM) with only one state, represented by a GMM with 8 components. The system completed with a Normalized Cross Likelihood Ratio (NCLR [35]) based on bottom-up clustering. Viterbi decoding is performed to adjust the segment boundaries. Non-speech regions are removed using GMMs. A hierarchical clustering for speaker models (using GMMs) is carried out over the clusters generated by the Viterbi decoding. So our system is trained to distinguish between primary child, secondary child, adult and non-speech classes instead of different speakers by MAP adapting the UBM to each of the three classes (child, primary adult, and secondary adult). The inputs to this system are 13 MFCCs with coefficient C0 as energy. In order to identify and remove music and jingle regions, the audio is segmented into speech and non-speech regions using a Viterbi decoding with 8 one state HMMs, comprising of 2 models of silence (wide and narrow band), 3 models of wide band speech (clean, over noise or over music), 1 model of narrow band speech, 1 model of jingles, and 1 model of music. Here, the features are 12 MFCCs completed by delta coefficients (coefficient C0 is removed).

4.3. Experimental results and discussion

In this section, we start by showing a confusion matrix for our i-vector SVM based diarization system, and then we present a case study after applying our child-adult diarization system to the utterances from 6 children with the childcare center. During the scoring, the mis-classification errors, are measured at a frame level by comparing the hypothesis classification with the reference segmentation generated by hand according to the audio content. Our experimental results show that applying TO-COMBO-SAD prior to i-vector based classification (section 4.1) results in up to 27.3% relative classification error rate reduction compared to the baseline system (section 4.2).

Table 2: *Confusion matrix for the i-vector SVM TO-COMBO SAD system with 1.5s segments*

I-vector SVM system	Error rate	Adult	Prim. child	Sec. child	Non-speach
Adult	13.6%	-	2.3	4.5	6.8
Prim. child	23%	4	-	7	12
Sec. child	26.2 %	4.4	8.7	-	13.12
Non-speech	35.6%	8.3	12.1	15.2	-

Table 2 shows the confusion matrix corresponding to our TO-COMBO-SAD based i-vector system using the data segments from the childcare center database. The segment length 1.5 seconds is chosen empirically. Comparing the individual class error rates (second column) across Table 2 shows that our system achieves lower error rates for adult and primary child classification and the error rates increase for non-speech and secondary child classification.

The non-speech confusion with other classes can be explained by the broad nature of its class (music, background noise, crowd noise, and singing). There is a considerable amount of background noise (including hints of child and adult distant speech) within our childcare center database, and this may be the reason behind the mis-recognition of child and adult speech segments as non-speech. For instance in our system 6.8%, 12%, and 13.12% of the errors occurred as a result of confusion between the adult, primary child, and secondary child classes and the non-speech class respectively (last column). The highest confusions occurred between the secondary child and non-speech classes. For instance, 13.12% of the classification errors in the secondary child classification (4th row), and 15.2% (last row) of the classification errors in the non-speech classification are due to the confusions between the secondary child speech and non-speech groups respectively. This might be due to the fact that a considerable amount of child speech from a distant proximity exists within the crowd noise (non-speech).

Using our diarization system (72% accuracy), we present a case study from 6 children, four males and two females, aged between 2.5 to 3 years old. One of these children has a developmental delay while the remaining are typically developing children. As shown in Table 3 one third of these children have been exposed to non-English primary language.

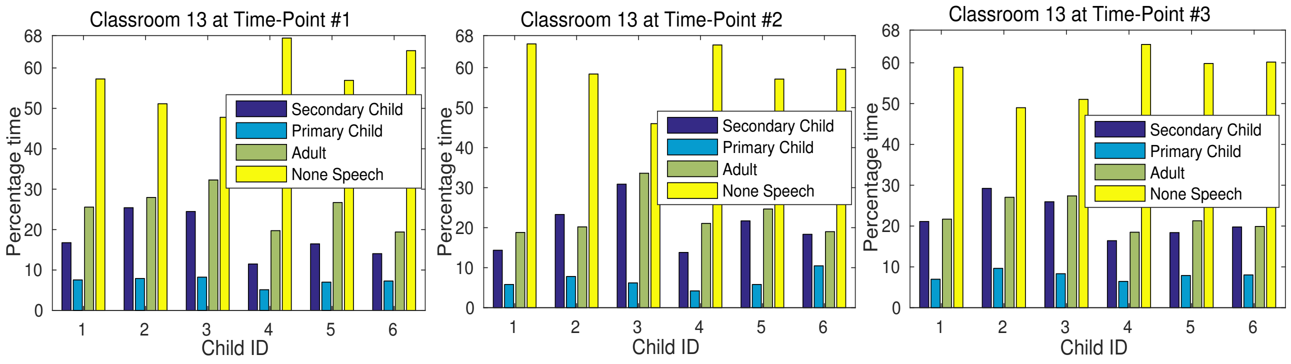
Table 3: *Child identity and further details*

Child ID	Age	Gender	Speech development	Primary language
1	3 yrs., 2 mths.	Male	typical	Turkish
2	3 yrs., 3 mths.	Male	delayed	English
3	3 yrs., 1 mths.	Female	typical	English
4	3 yrs., 2 mths.	Male	typical	Turkish
5	3 yrs., 1 mth.	Male	typical	English
6	3 yrs., 2 mths.	Female	typical	English

For each child, three audio recordings (3 hours each) was chosen from three typical days at three different Time-Points at the child care center (giving us a total of 9 hours of evaluation data per child). As shown in Figure 4, we aim to analyze the child's language environment by studying what percentage of time child is vocalizing in an activity that exposes him to a **non-speech** environment (yellow bars), versus what percentage of time the speech was produced by the **primary child** (blue bars), **secondary child** (purple bars), **adult** (green bars).

Here, the analysis is carried out during three distinct Time-Points. During Time-Point #1 and #3, Figures 4a and 4c respectively, audio files were recorded during the morning where the class schedule consisted of story book discussions, small group activities, free play, art activities, music activities, singing and dancing. During the Time-Point #2, Figure 4c,audio files were recorded during the afternoon where the class schedule consisted of free play, art activities, hand washing, having snacks, nap and quite time. It is interesting to compare the language environment across different children and compare the level of interaction for each individual child across Time-Points which involves different activities. Figure 4a shows that during Time-Point #1, the average duration of conversation directed by the teacher to the primary child reaches its maximum and the average duration of converstaion directed by other children to the primary child reaches a minimum compared to Time-Points #2 and #3. The relative higher level of conversation directed to child by the adult mainly occurs during story telling session were the teacher reads or explains something and asks the individual child's opinion on the topic. During such sessions, lower amount of speech will be directed to the child by other children. This amount shows that the nature of language environment that child is exposed to was more adult oriented compared to that of activities shown in Figures 4b and 4c for which this value is reaches to 4.20 and 3.8.

During Time-Point #1, on average the least amount of time has been spent for the conversation between child #4 and other teachers and children. This may be is because the child's primary language is not English and he needs more help to engage in conversation. On the other hand child #3 shows maximum interaction with other children and adults which might refer to



(a) (b) (c)

Figure 4: *A comparison among the level of interaction that each child had with other children and other adults*

the level of interest of this student in class. Despite the fact that child #2 had symptoms of development delays on average a relatively high amount of conversation was carried out between this child and other children and teachers. The yellow bar charts labeled non-speech refers to the average amount of audio in which there was no direct interaction between that child and other children and teachers. This includes the time spent during group singing, nap, lunch, play, or dance activities where the child was not talking to another child or adults directly. In the case of child #4 this graph shows that this child has spent a large amount of time during such activities.

Figure 4b shows that during Time-Point #2 which included the afternoon activities average activities that exposes the child to non-speech language environment reach a maximum compared to Time-Point #1 and #3 which are recorded during the morning. The large amount of non-speech duration can be explained by occurrences of nap and eating events during Time-Point #2. For child #4, similar to the Time-Point #1, on average there was a smaller amount of conversation taking place between this child and other children and the teachers. Across three Time-Points, the highest relative average amount of speech produced by primary and secondary children are taken place during Time-Point #3. This can be explained by the large amount of group activities and group play events at this Time-Point. This information can be useful for teachers and notifies them regarding the engagement and interaction of each child with peers and adults during the class activities compared to other children.

Collectively, the three Time-Points corresponding to 6 children in Figure 4 allow us to gain a wider perspective of child engagement with teachers in the classroom. From these automatic analysis plots we can begin to:

- Determine which children are less engaging in voice communication (i.e., child #4 very low vs. Child #3 more active).
- Determine how much engagement teachers have with each child (e.g., more with child #3, less with child #4 and #6).
- Assess on how much engagement each child has with other children during activities (e.g., child #2 and #3 more, child #4 and #6 less).
- Which events/activities stimulate greater voice communication between child-teacher and child-child.

5. Summary and future work

This paper has presented a close to real-time (1.5 seconds delay) adult-child speech diarization system with 72% accuracy which facilitates understanding of the child language environment.

Our experimental results show that applying TO-COMBO-SAD prior to i-vector based classification (section 4.1) results in up to 27.3% relative error rate reduction compared to the baseline results produced by the LIUM diarization toolkit (section 4.2).

Looking at the confusion matrix of the TO-COMBO-SAD i-vector based system (2) shows that our system achieves lower error rates for adult and primary child classification and the error rates increase for non-speech and secondary child classification. There is a considerable amount of background noise within our childcare center database, and this may be the reason behind the mis-recognition of child and adult speech as non-speech.

Using the proposed speech turn taking system we analyzed the child-child and teacher-child communication during a case study and compared them across different children, activities, and classes. This study showed the importance of speech analysis in building a foundation for automated child language environment analysis. These results can help us to measure the amount of verbalizations, both spoken and heard at different point of time, which may be sign of child's interest in specific activity. In addition to that, such evidence can also identify whether the amount of support and interaction received by each child from the teachers is a uniform across all children and across activity areas.

In the next stage of this research, we will investigate the on-line GMM-HMM based approach proposed in [36] to further improve our system's accuracy. In addition to that, we will apply subsequent analysis based on the word count [37], speech recognition [38, 39, 40], and keyword spotting [41]. For example determining word count and a list of most frequent keywords each child encounters can give us more insight into the richness of the child language environment. Studies have shown that rich language environments can contribute to essential developmental outcomes in early childhood [1].

Providing teachers with information about the language environment children experience and the locations they occupy will in all likelihood allow early educators to better orchestrate interactions in the classroom that support children's social and pre-academic learning.

6. References

[1] B. Hart and T. R. Riskey, "Meaningful differences in everyday experience of young american children," in *Hart and Risley*, 2004.

[2] *http://www.lenafoundation.org/.*

[3] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Prof-Life-Log: Personal interaction analysis for naturalistic audio streams," in *ICASSP*. IEEE, 2013, pp. 7770–7774.

[4] S. H. Yella, "Speaker diarization of spontaneous meeting room conversations," 2015.

[5] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *DSP*, vol. 10, no. 1, pp. 93–112, 2000.

[6] E. Gonina, G. Friedland, H. Cook, and K. Keutzer, "Fast speaker diarization using a high-level scripting language," in *ASRU*. IEEE, 2011, pp. 553–558.

[7] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *ASLP, IEEE Transactions on*, vol. 14, no. 5, pp. 1505–1512, 2006.

[8] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics." in *ICSLP*, 1998.

[9] J.-L. Gauvain and C.-H. Lee, "Bayesian learning of Gaussian mixture densities for hidden Markov models," *SNL*, vol. 2, p. 5, 1991.

[10] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, 1997.

[11] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[12] G. Gravier, M. Betser, and M. Ben, "AudioSeg: Audio segmentation toolkit, release 1.2," *IRISA*, 2010.

[13] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. W. Evans, B. G. Fauve, and J. S. Mason, "ALIZE/spkdet: a state-of-the-art open source software for speaker recognition." in *ODYSSEY*, 2008, p. 20.

[14] M. A. H. Huijbregts, "Segmentation, diarization and speech transcription: surprise data unraveled," in *Centre for Telematics and Information Technology University of Twente*, 2008.

[15] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *CSL*, vol. 20, no. 2, pp. 303–330, 2006.

[16] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings." in *INTERSPEECH*, 2012, pp. 2170–2173.

[17] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," IDIAP, Tech. Rep., 2013.

[18] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

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

[20] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen, and D. W. Oard, "Speech activity detection for NASA apollo space missions: challenges and solutions," in *INTERSPEECH*, 2014, pp. 1544–1548.

[21] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. L. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra *et al.*, "All for one: feature combination for highly channel-degraded speech activity detection." in *INTERSPEECH*, 2013, pp. 709–713.

[22] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 197–200, 2013.

[23] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM*, 2005.

[24] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *INTERSPEECH*, vol. 19, no. 4, pp. 788–798, 2011.

[25] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech." in *ODYSSEY*, 2010, p. 6.

[26] S. Safavi, M. Najafian, A. Hanani, M. Russell, and P. Jančovič, "Comparison of speaker verification performance for adult and child speech," in *WOCCI*, 2014.

[27] S. Safavi, M. Najafian, A. Hanani, M. J. Russell, P. Jancovic, and M. J. Carey, "Speaker recognition for children's speech," in *INTERSPEECH*, 2012, pp. 1836–1839.

[28] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH*, 2011, pp. 857–860.

[29] M. Najafian, S. Safavi, P. Weber, and M. Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *ODYSSEY*, 2016, pp. 213–218.

[30] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," *INTERSPEECH*, 2014.

[31] M. Najafian, S. Safavi, A. Hanani, and M. Russell, "Acoustic model selection for recognition if regional accented speech," *EUSIPCO*, 2014.

[32] M. Najafian, "Acoustic model selection for recognition of regional accented speech," Ph.D. dissertation, Ph. D. dissertation, University of Birmingham, 2016.

[33] S. Safavi, M. J. Russell, and P. Jancovic, "Identification of age-group from children's speech by computers and humans." in *INTERSPEECH*, 2014, pp. 243–247.

[34] M. Najafian, D. Irvin, Y. Luo, B. S. Rous, and J. H. L. Hansen, "Employing speech and location information for automatic assessment of child language environments," in *International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016, pp. 65–69.

[35] V. B. Le, O. Mella, D. Fohr *et al.*, "Speaker diarization using normalized cross likelihood ratio." in *INTERSPEECH*, vol. 7, 2007, pp. 1869–1872.

[36] R. Nabiei, M. Najafian, M. Parekh, P. Jancovic, and M. Russell, "Delay reduction in real-time recognition of human activity for stroke rehabilitation," *International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, pp. 70–74, 2016.

[37] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "A speech system for estimating daily word counts." in *INTERSPEECH*, 2014, pp. 880–884.

[38] S. Mirsamadi and J. H. L. Hansen, "A study on deep neural network acoustic model adaptation for robust far-field speech recognition," in *ISCA*, 2015.

[39] ——, "Multichannel feature enhancement in distributed microphone arrays for robust distant speech recognition in smart rooms," in *SLT*. IEEE, 2014, pp. 507–512.

[40] E. Fringi, J. F. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *ISCA*, 2015.

[41] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Automatic audio sentiment extraction using keyword spotting," in *ISCA*, 2015.