Tagging child-adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system

Prasanna V. Kothalkar[1], *Dwight Irvin*[2], *Ying Luo*[3], *Joanne Rojas*[3], *John Nash*[3], *Beth Rous*[3], *John H. L. Hansen*[1]

[1]Center for Robust Speech Systems(CRSS), University of Texas at Dallas, Richardson, TX, USA
[2]Juniper Garden's Children's Project, University of Kansas, KS, USA
[3]College of Education, University of Kentucky, KY, USA

prasanna.kothalkar@utdallas.edu, dwirvin@ku.edu, ying.luo@pnw.edu, joanne.rojas@uky.edu, john.nash@uky.edu, beth.rous@uky.edu, john.hansen@utdallas.edu

Abstract

Assessing child growth in terms of speech and language is a crucial indicator of long term learning ability and life-long progress. Since the preschool classroom provides a potent opportunity for monitoring growth in young children's interactions, analyzing such data has come into prominence for early childhood researchers. The foremost task of any analysis of such naturalistic recordings would involve parsing and tagging the interactions between adults and young children. An automated tagging system will provide child interaction metrics and would be important for any further processing. This study investigates the language environment of 3-5 year old children using a CRSS based diarization strategy employing an i-vector-based baseline that captures adult-to-child or child-to-child rapid conversational turns in a naturalistic noisy early childhood setting. We provide analysis of various loss functions and learning algorithms using Deep Neural Networks to separate child speech from adult speech. Performance is measured in terms of diarization error rate, Jaccard error rate and shows good results for tagging adult vs children's speech. Distinction between primary and secondary child would be useful for monitoring a given child and analysis is provided for the same. Our diarization system provides insights into the direction for preprocessing and analyzing challenging naturalistic daylong child speech recordings.

Index Terms: speech activity detection, child speech diarization, naturalistic environment, TO-Combo SAD, i-Vectors, Deep Neural Networks

1. Introduction

The diversity of language background, socio-economic conditions, development level, or potential communication disorders represents a challenge in assessment of child speech and language skills [1]. The language environment of young children plays an important role in development of speech, language, vocabulary and thus, thinking and learning ability, and has an impact on the life prospects of the child. The quality and number of interaction in a rich language environment helps in meeting essential language development outcomes in early childhood[2]. Thus, early childhood researchers are focusing on analyzing classroom interactions of preschool children to monitor and provide proactive support to them. Given the huge amount of daylong recordings to be analyzed, using automated speech processing and machine learning techniques would be highly beneficial. Previous classroom-based speech analysis systems have studied interaction of students in peer led team environment but for older children to provide communication metrics like word counts and speech qualities like curiosity, dominance, emphasis, engagement etc. The main challenge of such environments involve rapid short conversational turns, overlapped speech, noise and reverberation.

The preliminary task of analyzing such data environments involve speech diarization i.e. segmenting and tagging 'who spoke when'. Once this basic task is completed, further processing as mentioned above can be performed. In this study, we perform diarization on child-adult and child-child interactions of preschool children in naturalistic active learning environments. The audio data was collected using LENA devices[3, 4] worn by the children in different classrooms at different times. The recordings continue as subjects move around during a school day and are paused during nap time.

We provide baseline results using LIUM diarization toolkit[5] on gold standard segments. CRSS diarization toolkit is used for improving on the baseline. In this study, we present an i-Vector based Speaker Diarization system that performs segmentation using Speech Activity Detection and classification using a Deep Neural Network (DNN) model. Additionally, we compare different learning algorithms along with their loss functions to know the best performing configuration. Previous work on this dataset[6] used much lesser data and fixed segments of length 1.5 seconds with a Support Vector Machine (SVM) backend for classification. We have more data to improve the i-Vector training but utilize a challenging smaller segment length as we use i-Vectors as our features.

2. Data specifics

The dataset in this paper consists of spontaneous conversational speech recorded with the help of LENA units attached to the subjects in a high quality childcare learning center in the United States. The 48 recording sessions have children who are 3 to 5 year olds. About 15 hours (120K word tokens) of child speech was manually transcribed by the CRSS transciption team at UT Dallas. Another 23 hours of adult speech from 4 teachers/caregivers was manually transcribed providing 300K words in the transcripts. A total of 79 hours of speech and non-speech child and adult data was tagged by our transcribers. Three sessions have been excluded for further pre-processing/analysis, as we were unable to extract i-Vectors. We have divided the 45 session into training, development and test sets for training diarization system and evaluating its performance.

The training, development and test set divisions in terms of children, their sessions and total child-adult speech dura-

10.21437/SLaTE.2019-17

tions are shown in table 1. Gold standard speech segments are used for training and development while 0.5 second speech segments based on Threshold Optimized-ComboSAD(TO-ComboSAD)[7] evaluations are used in the testing stage. Although our SAD system provides highly accurate results, we can only select segments that can be assigned a reference speaker, given the segment start time and duration. This is due to fixed segment size of 0.5 seconds which eliminates segments that may have multiple speakers or overlapped speech and reduces the test evaluation data to 11 hours, 13 minutes and 52 seconds.

Set	**Child ID**	**Sessions**	**Duration (hh:mm:ss)**
Training Set	1	3	4:14:49
	2	2	3:22:52
	3	2	3:35:28
	4	2	2:48:42
	5	2	2:28:01
	6	1	1:26:55
	7	1	6:31:44
	8	1	2:14:07
	9	1	32:40
	10	1	2:10:49
Aggregate sum	**10**	**16**	**29:26:08**
Development Set	11	2	1:32:26
	12	2	2:38:07
	13	2	5:50:26
	14	1	54:44
	15	1	1:17:59
Aggregate sum	**5**	**8**	**12:13:42**
Testing Set	16	3	4:27:52
	17	3	1:43:47
	18	3	3:19:00
	19	2	2:37:39
	20	2	3:19:30
	21	2	5:17:19
	22	2	5:17:11
	23	2	8:23:47
	24	1	1:45:41
	25	1	1:01:19
Aggregate sum	**10**	**21**	**37:45:18**

Table 1: *Setwise Database Details*

3. Method

Our task is to tag quick conversational turns from the LENA audio data as being from Primary Child (PC), Secondary Children (SC) or Adults (AD). Here, PC carries the LENA recording device on his person, while SC/AD are the other children/adults that are recorded by the PC's LENA device.

3.1. Speech Activity Detection

Our toolkit TO-ComboSAD has performed extremely well with such long duration data [4, 7, 8]. TO-Combo SAD computes five noise robust features at the frame level for each segment and projects it into a single 1-dimensional space using Principal Component Analysis. The goal is to classify each audio file into speech and non-speech regions based on feature values at the frame level.

So in our case of daylong childcare center LENA unit recordings, each complete audio file is split into files of 20 minutes duration and then TO-ComboSAD is applied. This leads to better performance than entire 8-10 hours of audio recording, as there is sufficient data to train the model and computationally efficient. It trains a two-mixture GMM and finds the means for speech and non-speech regions. Let us denote the means by ν_{hs} for speech and ν_{hp} for background. The mixture with larger mean value is hypothesized to contain speech and vice-versa. This is from the fact that Combo features are designed to have higher values for speech and lower values for noise, background and silence.

Further, the mixture means are used to compute SAD thresholds which are used for speech/non-speech decisions.A large mixture Gaussian Mixture Model(GMM) model is learned from features extracted from annotated corpora of Switchboard and Fisher data. The m^{th} mean of the N mixture GMM is projected in the Combo-SAD dimension. Let the projected estimate of the mean be $\hat{\nu_n}$ and it's mean be ν_{ts}. Thus, ν_{ts} can be seen as the prior model of speech based on standard datasets and ν_{hs} will be the posterior model of the speech trained on the data. The threshold value α is computed as a convex combination of the estimated Gaussian means of the projected Combo features and is given by:

$$\alpha = k \max(\nu_{\mathrm{hs}}, \nu_{\mathrm{ts}}) + (1-k)(\nu_{hp}) \tag{1}$$

3.2. I-Vectors for Speaker Characteristics

I-Vectors [9] are fixed length vectors that characterize speaker identity from arbitrary length sequential data (i.e. speech samples). Factor analysis is performed to separate speaker dependent and speaker independent factors to represent unique attributes of the speaker. I-Vectors have been used for speaker recognition[10], language recognition[11], accent recognition[12], emotion recognition[13] etc. Within the child speech area, they have been used for speaker recognition [14, 15], age group identification [16] and screening children that can be 'at risk' of child speech disorders [17].

I-Vectors can be expressed by the following equation,

$$M = m + Tw \tag{2}$$

where m is the GMM supervector also known as Universal Background Model, T is the total variability matrix or i-Vector extractor, and w is the i-Vector. Here, T is the matrix of bases spanning the subspace for speaker and channel variability in the supervector space, and hence known as the total variability space, and w is standard normally distributed latent variable. For each observation sequence representing a speech utterance, our i-Vector is a Maximum-A-Posteriori (MAP) point estimate of the latent variable w.

In our system, 20-dimensional Mel-Frequency Cepstral Coefficients were extracted to model 256 gaussians and providing sufficient statistics for i-Vectors of 32 dimensions. The UBM model is trained on the development set and the TV matrix is trained on the training data.

3.3. DNNs for Speaker Classification

The block diagrams for training and development sets (Fig. 1) present our technique for developing the DNN Model, while block diagram for testing set (Fig. 2) presents our testing strategy using the trained DNN model. We used a Deep Neural Network (implemented in keras [18]) for classifying the labels as it

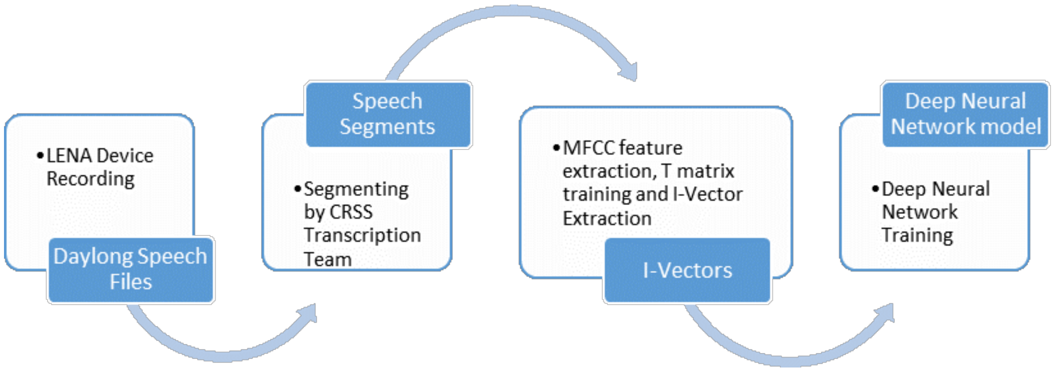
•LENA Device Recording
Daylong Speech Files
Speech Segments
•Segmenting by CRSS Transcription Team
•MFCC feature extraction, T matrix training and I-Vector Extraction
I-Vectors
Deep Neural Network model
•Deep Neural Network Training

Figure 1: *Block diagram for training and development data.*

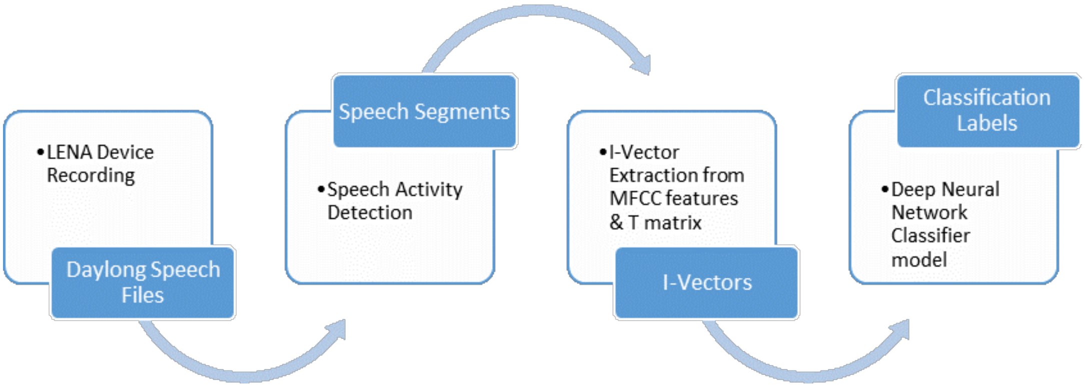
•LENA Device Recording
Daylong Speech Files
Speech Segments
•Speech Activity Detection
•I-Vector Extraction from MFCC features & T matrix
I-Vectors
Classification Labels
•Deep Neural Network Classifier model

Figure 2: *Block diagram for testing data.*

should be able to learn non-linear high level features for classification task. Previously i-Vectors have been used along with DNNs for classifcation tasks with good results[19, 20]. Another reason being, it provided high recall for PC versus multiple other classifiers including SVMs. This specific architecture first learns complex features in a high-dimensional space and then reduces number neurons in the last two layers, to provide the effect of dimensionality reduction. Various combination of layers, number of neurons, activation functions etc. were trained on training set and the current configuration was chosen based on performance on the validation set.

Parameter	**Value**
Number of layers	7
Batch size	128
Number of epochs	8
Input dimension	32 dimensions
Intermediate layer dimensions	[48,48,256,512,512,256,48]
Learning rate	0.002
Dropout rate	0.3
Layer numbers having droupout	[3,4,5,6]
Output dimensions	3 dimensions
Activation Function	Exponential Linear Unit
Loss Function	Categorical Cross Entropy (CCE), Mean Squared Error (MSE), Logcosh
Algorithms	Adadelta, Adam, Adamax

Table 2: *DNN Architecture Experimental Setup*

4. Results and Discussion

4.1. Equal Error Rate

Equal Error Rate can be defined as the error rate at the point of the detection error tradeoff curve where the False Alarm Rate is equal to the False Rejection Rate. For the task of Speech Activity Detection, our test dataset provides an average Equal Error Rate of 1.35%.

4.2. Diarization Error Rate

Diarization results are presented in terms of diarization error rate (DER) that can be defined as the sum of errors due to incorrect speaker ($\mathrm{E_{spkr}}$), missed speech ($\mathrm{E_{MISS}}$), false alarm speech ($\mathrm{E_{FA}}$) and overlapping speakers ($\mathrm{E_{ovl}}$).

$\mathrm{E_{spkr}}$: Percentage of scored time that a speaker ID is assigned to the wrong speaker.

$\mathrm{E_{MISS}}$: Percentage of scored time that a hypothesized non-speech segment corresponds to a speaker segment.

$\mathrm{E_{FA}}$: Percentage of scored time that a hypothesized speaker segment is labelled as non-speech in the reference.

$\mathrm{E_{ovl}}$: Percentage of scored time that some of the multiple speakers in a segment do not get assigned to any speaker.

$$\mathrm{DER} = \mathrm{E_{spkr}} + \mathrm{E_{MISS}} + \mathrm{E_{FA}} + \mathrm{E_{ovl}} \tag{3}$$

Applying LIUM diarization toolkit[5] to gold standard test speech segments provided an average DER in excess of 100%. So we did not try more experiments of fixed length segments.

4.3. Jaccard Error Rate

We also report Jaccard Error Rate (JER), a metric introduced for DIHARD II[21] that is based on the Jaccard index. The Jaccard index is a similarity measure typically used to evaluate the output of image segmentation systems and is defined as the ratio between the intersection and union of two segmentations. To compute Jaccard error rate, an optimal mapping between reference and system speakers is determined and for each pair the Jaccard index of their segmentations is computed. The Jaccard error rate is then 1 minus the average of these scores.

An optimal mapping between speakers is determined using the Hungarian algorithm so that each reference speaker is paired with at most one system speaker and each system speaker with at most one reference speaker. Then, for each reference speaker, the speaker-specific Jaccard error rate is $\frac{(\mathrm{E}_{\mathrm{FA}} + \mathrm{E}_{\mathrm{MISS}})}{\mathrm{E}_{\mathrm{TOTAL}}}$.

$\mathrm{E}_{\mathrm{TOTAL}}$: The duration of the union of reference and system speaker segments; if the reference speaker was not paired with a system speaker, it is the duration of all reference speaker segments.

E_{FA}: The total system speaker time not attributed to the reference speaker; if the reference speaker was not paired with a system speaker, it is 0.

$\mathrm{E}_{\mathrm{MISS}}$: The total reference speaker time not attributed to the system speaker; if the reference speaker was not paired with a system speaker, it is equal to TOTAL.

The Jaccard error rate then is the average of the speaker specific Jaccard error rates. Results for current system in terms of DER and JER are presented in Table 3.

Algorithm	Loss function	DER	JER
Adadelta	MSE	39.1%	63.9%
Adam	MSE	39.9%	66.4%
Adamax	MSE	39.9%	63.7%
Adadelta	Logcosh	38.1%	64.2%
Adam	Logcosh	39.5%	64.1%
Adamax	Logcosh	41.8%	65.2%
Adadelta	CCE	40.8%	62.5%
Adam	CCE	40.2%	70.5%
Adamax	CCE	**37.3%**	**62.0%**

Table 3: *Diarization error rate and Jaccard error rate diarization results*

The best results are provided by adamax algorithm with categorical cross entropy loss function. Table 4 shows the confusion matrix for the three classes (in terms of accuracy), we aim to predict for the best (algorithm-loss function) combination. Accuracy can be defined as the fraction of samples that belong to a class and have been predicted correctly. Our unweighted average recall (50.2%) is better than chance (33.3%) and the adults' class shows the highest accuracy followed by secondary children and primary child. This signfies that i-Vectors are able to capture adult patterns in the data correctly. Around half of the PC and SC speech segments are predicted as adults. This means better feature modelling techniques would be desired for recognizing children's speech using i-Vectors. This could also be due to the data imbalance working in favor of AD (49,854 utterances) > SC (18,040 utterances) and AD > PC (20,799 utterances). Additionaly, PC (24.23%) is harder to predict than SC (40.73%). Thus, class balanced metrics could provide complementary information while evaluating system performance and data imbalance should be taken into account while modelling future systems. Also, 26.86% of the PC data was recognized as SC, while only 9.07% of SC data was tagged as PC, despite there being more PC utterances. Thus, better strategies need to be devised for separating PC and SC.

		System		
		PC	SC	AD
Reference	PC	**24.2%**	26.9%	48.9%
	SC	9.1%	**40.7%**	50.2%
	AD	4.3%	10.2%	**85.5%**

Table 4: *Confusion Matrix for Primary Child, Secondary Children and Adults using Adamax algorithm and Categorical Cross Entropy Loss function*

5. Conclusions and Future work

This study presents a preliminary system for diarization of day-long child speech recordings in a child care learning environment. A combination of I-Vector and DNN Classification system provides effective diarization error rate and Jaccard error rate. The complete pre-processing pipeline includes excellent Speech Activity Detection followed by tagging 0.5 seconds speech segments. Further work towards an improved diarization system could be the initial step for multiple Child Speech Processing systems including Automatic Speech Recognition, Word Counting and Vocabulary Diversity measurement systems, screening kids with Speech Sound Disorders etc. Signal processing enhancement techniques, data augmentation, DNN embeddings and advanced DNN architectures can provide better results in terms of DER. Distance from microphone can be utilized for better recognition of PC from SC in future. Features enhancing child-specific speech characteristics would be helpful in improving adult versus child classifications. Also we would like to reduce the segment lengths to 0.25 seconds and also test with different segment durations.

6. Acknowledgements

The authors would like to thank all the volunteering participants.

7. References

[1] S. Rosenbaum and P. Simon, *Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program.* ERIC, 2016.

[2] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, 1995.

[3] "https://www.lenafoundation.org."

[4] A. Ziaei, A. Sangwan, and J. H. Hansen, "Prof-life-log: Personal interaction analysis for naturalistic audio streams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7770–7774.

[5] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[6] M. Najafian, D. Irvin, Y. Luo, B. S. Rous, and J. H. Hansen, "Automatic measurement and analysis of the child verbal communication using classroom acoustics within a child care center." in *WOCCI*, 2016, pp. 56–61.

[7] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech activity detection for NASA apollo space missions: Challenges and solutions," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] H. Dubey, L. Kaushik, A. Sangwan, and J. H. Hansen, "A speaker diarization system for studying peer-led team learning groups," *arXiv preprint arXiv:1606.07136*, 2016.

[9] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[11] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[12] M. H. Bahari, R. Saeidi, D. Van Leeuwen *et al.*, "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7344–7348.

[13] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[14] S. Safavi, M. Najafian, A. Hanani, M. J. Russell, and P. Jancovic, "Comparison of speaker verification performance for adult and child speech." in *WOCCI*, 2014, pp. 27–31.

[15] S. Safavi, M. Najafian, A. Hanani, M. J. Russell, P. Jancovic, and M. J. Carey, "Speaker recognition for children's speech," *arXiv preprint arXiv:1609.07498*, 2016.

[16] S. Safavi, M. Russell, and P. Jančovič, "Identification of age-group from children's speech by computers and humans," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[17] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell, and J. H. Hansen, "Fusing text-dependent word-level i-vector models to screen'at risk'child speech," *Age (months)*, vol. 51, no. 11, pp. 36–78, 2018.

[18] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[19] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples." in *Interspeech*, 2016, pp. 1195–1199.

[20] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[21] "https://coml.lscp.ens.fr/dihard/index.html."