International Journal of Speech Technology
https://doi.org/10.1007/s10772-019-09590-0

Speech and language processing for assessing child–adult interaction based on diarization and location

John H. L. Hansen[1] · **Maryam Najafian**[1] · **Rasa Lileikyte**[1] · **Dwight Irvin**[2,3] · **Beth Rous**[3]

Received: 9 August 2018 / Accepted: 9 January 2019

Abstract
Understanding and assessing child verbal communication patterns is critical in facilitating effective language development. Typically speaker diarization is performed to explore children's verbal engagement. Understanding which activity areas stimulate verbal communication can help promote more efficient language development. In this study, we present a two-stage children vocal engagement prediction system that consists of (1) a near to real-time, noise robust system that measures the duration of child-to-adult and child-to-child conversations, and tracks the number of conversational turn-takings, (2) a novel child location tracking strategy, that determines in which activity areas a child spends most/least of their time. A proposed child–adult turn-taking solution relies exclusively on vocal cues observed during the interaction between a child and other children, and/or classroom teachers. By employing a threshold optimized speech activity detection using a linear combination of voicing measures, it is possible to achieve effective speech/non-speech segment detection prior to conversion assessment. This TO-COMBO-SAD reduces classification error rates for adult-child audio by 21.34% and 27.3% compared to a baseline i-Vector and standard Bayesian Information Criterion diarization systems, respectively. In addition, this study presents a unique location tracking system adult-child that helps determine the quantity of child–adult communication in specific activity areas, and which activities stimulate voice communication engagement in a child–adult education space. We observe that our proposed location tracking solution offers unique opportunities to assess speech and language interaction for children, and quantify the location context which would contribute to improve verbal communication.

Keywords Child speech · Speaker diarization · Speech activity detection · I-Vector · Language environment monitoring

1 Introduction

Speaking and listening are primary communication modes in most educational settings. The language environment in early childhood is linked to children's language development. Additionally, a rich communicative experience early in childhood is essential for school readiness, early literacy and academic performance (Hart and Risley 1995; Walker et al. 1994). However, to our knowledge, there are no studies performed to establish a relation between children verbal communication quantity and activity/learning areas. Understanding which activity areas stimulate verbal communication can assist in developing improved context spaces/stations and thereby contribute to more efficient children language development.

For humans, analyzing a 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 systems that can bring consistency and reliability to the analysis. Access to automatic language environment monitoring systems can assist researchers and educators to objectively interpret measures of the amount of the childs engagement in speech communication (speech produced by the child or directed to the child) with respect to a target condition, and furthermore identify whether a child requires

✉ John H. L. Hansen
John.Hansen@utdallas.edu

[1] Center for Robust Speech Systems, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA

[2] Life Span Institute University of Kansas, Kansas City, KS, USA

[3] College of Education, University of Kentucky, Lexington, KY, USA

Published online: 05 June 2019

further assistance based on their vocal communication patterns (Hart and Risley 1995; Najafian et al. 2016; Gupta et al. 2016).

This study centers on two main issues that must be addressed to develop effective children vocal engagement assessment solutions:

1.1 Advanced child–adult diarization system

We develop an advanced near to real-time, noise robust system that can measure the duration of child-to-adult and child-to-child conversations, and can track the number of conversational turn-takings. It allows one to explore the amount of child engagement in conversations and determines how much of the childs interaction involves other children versus classroom teachers (Najafian et al. 2016). A speech activity detector followed by an i-Vector based child–adult turn-taking detection solution is developed. The advanced i-Vector based classification system was inspired by the success of i-Vector based systems in speaker recognition, and child age classification systems, but is designed to exploit child adult turn-takings with much smaller duration speech segments. The LENA[1] recording device (Ziaei et al. 2013) is employed and robust analytical algorithms for a machine-based solution are used.

1.2 Location tracking

In this study a child location tracking method is proposed, that to the best our knowledge, is a novel approach to determine a childs communication activity/learning areas and relate this with the quantity of children verbal communication. Location tracking consists of three components: (1) children's time spent in different language environments, (2) adult's vocal interaction level in these activity areas, (3) child's vocal interaction quantity across time points. From the results it can be observed where a child spends more/less of their time during activities, as well as the amount of verbalizations, both spoken and heard. Moreover, this information is helpful for educators to quantify interactions in the classroom which is essential for supporting a childs social and pre-academic learning.

For our experiments, we record and track the location of 33 children of age 2.5 to 5 years in age across 4 classrooms in a high-quality child care center in the United States at various time points during the day. Each classroom setting has between 14-20 children with typically two adult educators. Each LENA unit typically captures two way voise conversations between primary and secondary speakers.

[1] http://www.lenafoundation.org.

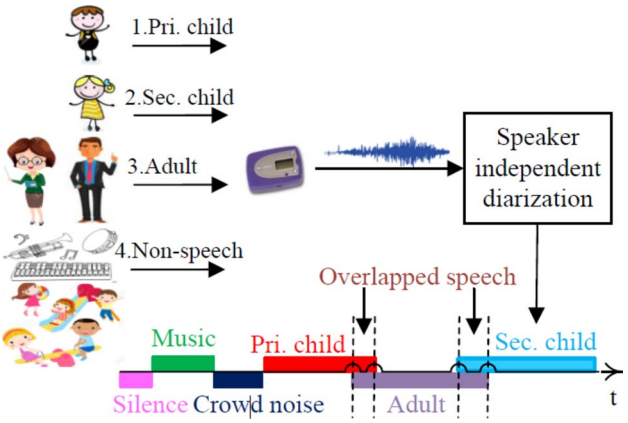

Fig. 1 General diarization system

The reminder of this paper is organized as follows: In Sect. 2 the challenges in diarization for child-adult scenario are highlighted. Section 3 describes the child–adult data set. In Sect. 4 we review related systems. We present details of our diarization system in Sect. 5, and provide experimental results in Sect. 5.4. In Sect. 6 our novel location tracking approach is described. Conclusions are drawn in Sect. 7.

2 Challenges in diarization for child–adult scenario

Childrens speech diarization within a naturalistic education space is a more challenging task than traditional adults speech diarization for a range of reasons as follows:

- Acoustic and linguistic characteristics of children's speech.
- Childs vocal system is smaller than an adult's, which changing size and shape over development. This leads to higher variability of child speakers and is more difficult to separate male and female children speakers.
- Child speakers have a higher variability in speaking rate.
- Children generally display a a higher degree of spontaneous speech, which may be ill-formed, or incomplete sentences.
- Paralinguistic evens are more common for children, such as non-speech vocalizations as laughter, crying, shouting, yawning, coughing, or sneezing.
- Due to a lack of social/communication experience, children speech will have a higher degree of overlap in naturalistic conversation interactions.
- Children are more likely to seamlessly change conversational topics rapidly.

Due to these features, naturalistic child child speech diarization is significally more challenging than traditional

diarization of single or two-speaker adult speech in telephony or broadcast news scenarios.

The proposed general diarization system employed in this study is illustrated in Fig. 1. Children speech tend to be significantly overlapping, because typically children lack social communication skills. So, while the adult may follow traditional conversational turn-taking protocol, children will speak with little regard to expected sentence bounders. With the such ill-formed spontaneous speech, traditional mentory of performance such as when become almost meaningless for this scenario. Therefore, other measures of conversational assessment may be necessary.

3 Data set

For speech data collection, a light weight compact digital audio recorder LENA device (Ziaei et al. 2013) is worn (e.g., for the data used in these experiments, this included 33 children of age 2.5 to 5 years old). The LENA unit capture as much as 16 h of continuesly recorded audio within a day, thought total recording per child here is typically 4–7 h. The audio is recorded throughout a typical day at an education/childcare center, at three time periods where the child was participating in different education/social activities. We used 4.5 h of audio recording gathered by the LENA unit attached to 18 children (approximately 15 min each) to train our speech analysis systems. In our experiments, a threefold cross validation scenario was used, so no speaker appeared simultaneously in the training and test sets. For system evaluation, data was partitioned into labeled segments. The labels identify whether the segment belongs to the following categories.

- Non-speech: the stream of background noise, silence, music or conversation produced by other children or adults who are generally more than 8 feet away from the primary child speaker.
- Speech:
  - Primary child: speech initiated by the child wearing the LENA unit.
  - Secondary child: speech originated by other children and directed at the primary child within his/her close proximity.
  - Adult: speech originated by a close adult and directed at the primary child within his/her close proximity.

From the manual human transcribed labels gathered, it was estimated 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 (see Fig. 2).

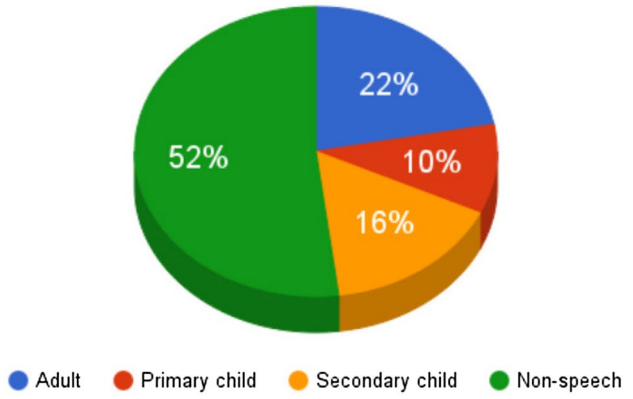

Fig. 2 Percentage of hand-labeled four way classes of data in the database, namely adults, primary child, secondary child, and non-speech

Table 1 Ground-truth analysis for the dataset

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

Table 1 reports average segment and turn durations within the database across primary child, secondary child and adult. The average segment duration for each class 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.

For location data collection purposes a real-time location tracking system Ubisense device (Woźniak et al. 2013) was worn by all participating children. Ubisense relies on receivers and transmitters, which communicate using ultra wide band radio frequencies to report the childs location every second. These communications are logged by PC running the Ubisense Location Engine software packages (Woźniak et al. 2013). With proper calibration, the accuracy of Ubisense is $\pm$ 15 cm under ideal measurement conditions, and $\pm$ 30 cm in challenging measurement conditions (Phebey 2010). Ubisense has been used in a variety of commercial and research endeavors (Swedberg 2011; Riehle et al. 2008; Connaghan et al. 2009).

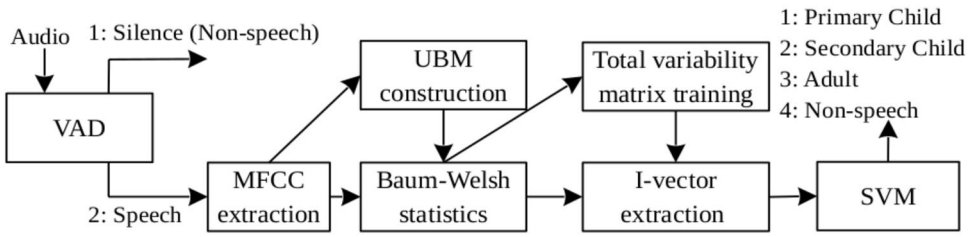

Fig. 3 i-Vector based child–adult event classification system

4 Related systems

Our child–adult turn-taking detection system has some similarities with the speech diarization systems, since it requires detecting turn change points as the source of the audio segment changes. In this section, the state of the in this area is revised, and in the next section it will be compared the performance of our advanced system with a state-of-the-art diarization system for the child–adult turn-taking tracking. 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) (Barras et al. 2006) is employed both as similarity measure and as stop criterion. Next, 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) (Reynolds et al. 1998) measure and GMMs trained with the Maximum A Posteriori algorithm (MAP) (Gauvain and Lee 1991). Using this diarization routine, several broadcast news and meeting diarization toolkits have proposed in the literature, namely the CMU Segmentation tool Siegler et al. (1997), the LIUM open-source speaker diarization toolbox (Meignier and Merlin 2010), the AudioSeg Audio segmentation toolkit (Gravier et al. 2010), the speaker diarization and recognition library ALIZE (Bonastre et al. 2008), the SHoUT diarization toolkit (Huijbregts 2008), the diarization system by LIA and CLIPS laboratories (Meignier et al. 2006), the IDIAP DiarTK toolkit (Vijayasenan and Valente 2012) where clustering and segmentation are based on the information bottleneck principle, and finally the recent work by Yella (2015) based on Information Bottleneck with Side Information (IBSI) which suppresses artifacts of background noise and non-speech segments at the conversation clustering phase. These systems can perform better when there are pauses between conversational turn-takings rather than spontaneous speech.

One of the main issues with most of these speaker diarization systems is the lack of a simple approach that can robustly and efficiently be applied to audio segments without the need for expensive agglomerative cluster merging and retraining, parameter tuning or adjusting minimum duration constraints for Viterbi realignment (Anguera et al. 2012; Tranter and Reynolds 2006). Previously the problem of clustering efficiency for large sets of speaker segments has been addressed by employing complete-linkage clustering (Ghaemmaghami et al. 2011), however the use of Viterbi realignment in their diarization module has resulted in inefficiencies when processing long recordings. A cluster-voting approach has been proposed in (Ghaemmaghami et al. 2015) which took advantage of multiple clustering decisions in order to make a more informed clustering decision without requiring Viterbi realignment to rectify incorrect clustering decisions.

5 Child–adult diarization system

One of the main issues with most of the speaker analysis and diarization systems is the lack of a simple approach that can robustly and efficiently be applied to audio segments without the need for expensive agglomerative cluster merging and retraining, parameter tuning or adjusting minimum duration constraints for Viterbi realignment. In this section our baseline system is decribed, that is an i-Vector based child–adult diarization with a Support Vector Machine (SVM) (Cortes and Vapnik 1995) (Sect. 5.1), then we modify this system by exploiting a Threshold Optimized Speech Activity Detector (TO-COMBO-SAD) (Ziaei et al. 2014) for separating the speech/non-speech segments at the beginning (Sect. 5.2). Finally, a speaker diarization system is described, that has been successfuly used previously in the broadcast news diarization task (Meignier and Merlin 2010) (Sect. 5.3).

5.1 Baseline system

In this section the description is provided of our baseline i-Vector based child–adult turn-taking detection system with 35.6% classification error rate. The system is illustrated in

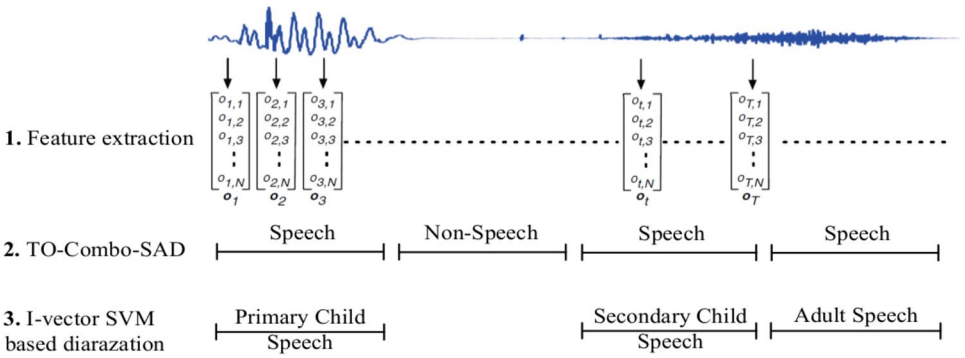

Fig. 4 Main stages of the child–adult turn-taking system

Fig. 3. The i-Vectors (Dehak et al. 2011a; Bahari et al. 2014) can best describe the coordinates of the audio features in a low dimensional space. The i-Vectors were successfully applied to, namely speaker recognition (Dehak et al. 2011b) language, and accent recognition areas. The advanced child–adult turn-taking detection system was inspired by the success of i-Vectors in age-group identification task for children and adults (Safavi et al. 2014). During the i-Vector approach only one single space (total variability space) is defined for describing all types of both speaker and session variabilities in an utterance as described below. In order to be able to capture rapid child–adult conversational turns while capturing useful information about the speaker in this study we segmented the audio recordings into 1.5 s cuts. In our system the rank of the i-Vector space is quite small (e.g., 25) compared with the number used in speaker recognition (e.g., 300) (Dehak et al. 2011a) or accent/language recognition (e.g., 500) due to the short estimation window.

5.1.1 Feature extraction and voice activity detection (VAD)

The speech is 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 0th to 12th-order Cepstral coefficients, log energy, and all delta and delta-delta variants.

5.1.2 UBM

Speech from the training set is used to estimate the parameters of the Universal Background Model (UBM).

5.1.3 Baum-Welch statistics

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

5.1.4 Extracting the i-Vectors

For child or adult utterances, the value of T-matrix and i-Vector (mean of posterior distribution) are estimated iteratively using the EM algorithm. In the Expectation step, T is assumed to be known, and w is updated. In the Maximization step, w is assumed to be known and T is updated. For utterance, u, in the Expectation step, the i-Vector w is updated using the current value of the T-matrix, and the Baum-Welch statistics extracted from the UBM. In our system, the UBM was trained on the training subset of the dataset using various number of UBM components and T-matrix ranks. Our system uses a UBM with 256 components, and a T-matrix of rank 25 (chosen empirically).

5.1.5 SVM

A multi-class SVM classifier with linear Kernel is trained to classify the i-Vectors into 3 classes. The test speakers 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 segment.

5.2 System advancements

In this section our threshold optimized i-Vector based child–adult turn-taking detection system with 28% classification error rate is presented. As shown in Fig. 4, the advanced system classifies the given audio into four main

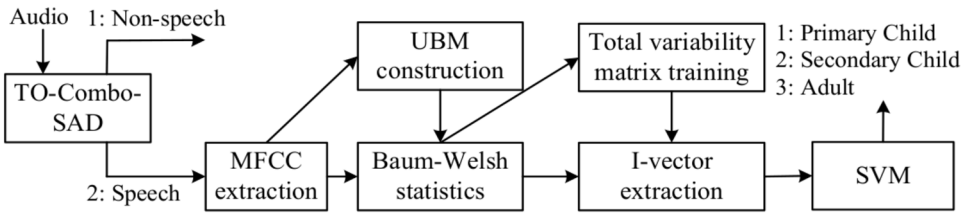

Fig. 5 Applying TO-COMBO-SAD prior to event classification

categories of non-speech, adult (teacher) speech, primary and secondary child speech using a Support Vector Machine (SVM) classifier (Cortes and Vapnik 1995). Our system uses Speech Activity Detection (SAD). This system starts by removing the non-speech regions (environmental noise such as crowd and music noise) using a TO-COMBO-SAD Ziaei et al. (2014) (Fig. 5). This speech activity detection has been particularly effective in multiple RATS evaluations (Sadjadi and Hansen 2013; Graciarena et al. 2013). During SAD several noise robust features are computed at a frame level for each audio segment and the combined feature vectors are projected into a single dimension (by using Principal Component Analysis) for the speech and non-speech discrimination task. This feature is efficiently obtained from a linear combination of the voicing measures, namely harmonicity, clarity, prediction gain, and periodicity as described bellow. After applying SAD, the previously described i-Vector based child–adult turn-taking detection system is applied to the speech segments.

5.2.1 Feature extraction

The audio signal is blocked into 25 ms frames (10 ms overlap). To extract the periodicity, harmonicity, and clarity, we choose a pitch period duration within the interval of (Delano and Snell 2006; Barras et al. 2006) ms (or equivalently [62.5, 500] Hz in the frequency domain), where the lower limit is imposed by the analysis frame length, and the fact that each frame should at least cover two pitch periods for a reliable voicing estimate.

5.2.2 Normalized autocorrelation estimation

Estimation of time domain voicing measures, such as harmonicity, clarity and prediction gain rely on normalized autocorrelation value proposed in (Boersma 1993). It has been shown that normalization by autocorrelation of the window function effectively mitigates the impact of strong formants on the maximum autocorrelation peak in the pitch range, obviates the need for low-pass filtering and/or center-clipping and compensates for the windowing effect. For noise robust pitch estimation, the deterministic autocorrelation of a short-time windowed segment needs to be computed.

5.2.3 Harmonicity feature

Harmonicity is defined as the relative height of the maximum autocorrelation peak in the plausible pitch range For voiced segments the harmonicity shows sharp peaks. Clarity feature: clarity is defined as the relative depth of the minimum average magnitude difference function in the plausible pitch range. The clarity exhibits large values for voiced and speech-like segments, while maintaining a minimum for background sounds.

5.2.4 Prediction gain feature

The prediction gain is defined as the ratio of the signal energy to the linear prediction residual signal energy. From the Levinson–Durbin recursion intermediate set of parameters is obtained that can be equated to the reflection coefficients of an acoustic tube model of the vocal tract.

5.2.5 Periodicity feature

The periodicity can thus be used to effectively discriminate speech from non-speech sounds. In the Short Time Fourier Transform (STFT) domain, the harmonics of the pitch frequency are apparent in the magnitude spectrum of speech during voiced and speech-like segments. This observation serves as the basis for the harmonic product spectrum (HPS) technique which has been widely applied for pitch detection in noisy environments. The periodicity is especially impervious to noise and other background sounds, since their spectral harmonics cannot combine coherently in the HPS. The frequency-compressed copies coincide at the fundamental frequency and re-inforce the amplitude, while other harmonics are attenuated in the final product. The periodicity is computed as in the plausible pitch range.

5.2.6 Spectral flux (SF) feature

The SF (Scheirer and Slaney 1997), is a feature capable of measuring the degree of variation in the spectrum across time. The negative of perceptual SF exhibits small values for non-speech segments (background sounds/silence),

Table 2 Confusion matrix for the i-Vector SVM system with 1.5s segments

i-Vector SVM system	Error rate (%)	Adult	Prim.child	Sec. child	Non-speech
Adult	19.5	–	3.2	7.3	9.1
Prim. child	34	6	–	9	19
Sec. child	35.6	5.6	10.6	–	19.37
Non-speech	42.7	10.8	15	16.9	–

while maintaining a maximum value for speech segments. the speech segments.

5.2.7 Principal component analysis (PCA)

A 5-dimensional vector is formed by concatenating the above named features. Each feature dimension is normalized (zero mean unit variance).The normalized feature vectors are linearly mapped into a 1-dimensional feature space using the PCA algorithm (Zhao et al. 1988). The 1-dimensional combo feature is smoothed via a 3-point median filter to serve as soft-decisions for the SAD (referred as Combo-SAD feature).

5.2.8 SAD threshold estimation

The Combo-SAD feature has a bimodal distribution in which speech and non-speech classes are well separated. In this step, the mixture means are used to compute the SAD threshold and speech/pause decisions are made. We exploit this property by fitting a 2-mixture GMM to the feature and estimating a detection threshold, θ from a weighted average of the mixture means, where α, μ_{hs} and μ_{hp} are the weight factor, hypothesized speech and non-speech mixture means, respectively. Next, the means of this GMM are projected into the Combo SADs single-dimension decision making space, where $\hat{m}_j$ is the j mixture mean of the M-mixture GMM, and m_j is the corresponding projected value. 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 based on Combo-SAD features from data). If $\mu_{hs} \geq \mu_{ts}$ then we trust the posterior model of speech and use it for decision making. Alternatively, if $\mu_{hs} \geq \mu_{ts}$, then we use the prior model of speech for decision making. Then speech/pause decisions are made using the SAD threshold value τ is based on a simple convex combination. During both training and testing of the child–adult turn-taking detection system only the segments labeled as speech are given as inputs.

5.3 LIUM speaker diarization toolkit

In this section we describe the application of the LIUM speaker diarization system (Meignier and Merlin 2010) for child–adult and non-speech classification with 38.5% error rate. This diarization process can break down into three main stages.

5.3.1 Feature extraction

The inputs to this system are 13 MFCCs with coefficient C0 as energy, computed every 10 ms using a 20 ms window.

5.3.2 Segmentation based on BIC

The initial segment boundaries are determined according to a Generalized Likelihood Ratio (GLR), computed using Gaussians with full covariance matrices. The Gaussians are estimated over a 2 s window sliding along the whole signal. A segment boundary (i.e. change point), is present in the middle of the window when the GLR reaches a local maximum.

5.3.3 BIC Clustering

The Universal Background Model (UBM) is adapted (Maximum A Posteriori) MAP for each cluster. The clustering is based on a bottom-up hierarchical agglomerative clustering (Siegler et al. 1997). In the initial set of clusters, each segment is a cluster. The two closest clusters are then merged at each iteration until the BIC stop criterion is satisfied. Each cluster, is modeled by a full covariance Gaussian during the segmentation process. The BIC penalty factor is computed over the length of the two candidate clusters.

5.3.4 Segmentation based on Viterbi decoding

A Viterbi decoding is performed to adjust segment boundaries. A cluster is modeled by a Hidden Markov Model (HMM) with only one state, represented by a GMM with 8 components learned by maximum-likelihood expectation maximization over the set of class label segments. A hierarchical clustering for speaker models (using GMMs) is carried out over the clusters generated by the Viterbi decoding.

5.3.5 Speech detection

Our system is trained to distinguish between primary child, secondary child, adult and non speech classes. In order to identify and remove music regions, the audio is segmented

Table 3 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-speech
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	–

crowd noise, singing). There is a considerable amount of background noise (including far distance child and adult speech) within our childcare center database, and this may be the reason behind the mis-recognition of child and adult speech segments as a non-speech group. For instance in the baseline system 9.1%, 19%, and 19.37%, and in the TO-COMBO-SAD based system 6.8%, 12%, and 13.12% of the errors occurred as a result of confusion between the, namely adult, primary child, and secondary child classes and the non-speech class respectively (last column).

Table 4 Effect of segment duration on the classification error

Child–adult turn-taking detection	Error rate	Error rate	Error rate	Error rate
Segments duration	1 s	1.5 s	2 s	3 s
TO-COMBO-SAD i-Vector SVM	32.1%	28 %	30.3%	31%
T-Matrix rank	20	25	150	200

into speech and non-speech regions using a Viterbi decoding with 8 one-state-HMMs, comprising of 1 model of silence, 1 model of background crowd noise, 1 model of music, 3 models of speech (primary child, secondary child, adult speech).

5.4 Results and analysis

In this section, firstly a confusion matrix for our i-Vector SVM based diarization system is showed. During the scoring the misclassification 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 an i-Vector based classification (Sect. 5.2) results in up to 21.34% and 27.3% relative classification error rate reduction compared to the baseline system (Sect. 5.1) and standard speaker diarization system (Sect. 5.3) respectively.

Table 2 shows the confusion matrix corresponding to our baseline i-Vector system, using the data segments from the childcare center database. The segment length 1.5 seconds is chosen impractically. Comparing the individual class error rates (second column) across Table 2, it can be seen 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.

Comparing the individual class error rates (second column) across Tables 3 and 2, it is observed that using the TO-COMBO-SAD prior to the i-Vector based baseline classification has resulted in up to 22% error rate reduction within each individual class. In both systems the lowest error rate belongs to the adult classification, and the highest error rates belong to the non-speech classification. The non-speech confusion with other classes can be explained by the broad nature of its class (music, background noise,

Table 5 Child identity and background information

Child ID	Age	Gender	Speech development	Primary language
1	3 years, 2 months	Male	Typical	Turkish
2	3 years, 3 months	Male	Delayed	English
3	3 years, 1 months	Female	Typical	English
4	3 years, 2 months	Male	Typical	Turkish
5	3 years, 2 months	Female	Typical	English

The highest confusions occurred between the secondary child and non-speech classes. For instance, in the baseline and the TO-COMBO-SAD based systems 19.37% and 13.12% of the classification errors in the secondary child classification (4th row) and 16.9% 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).

Table 4 presents the effect of segment duration on classification error rate for the TO-COMBO-SAD i-Vector SVM system. The lowest error rate is achieved when using segment duration of 1.5s. Previous research has shown that longer audio segments will result in capturing more speaker dependent information in the i-Vector space. On the other hand, selecting longer audio segments will result in missed turn-taking points and increasing the classification error rates. This is also evident from our results. For instance for audio segments of length 1 s (first column, Table 4) enables the system to capture rapid turn-takings, however this limits the amount of useful information

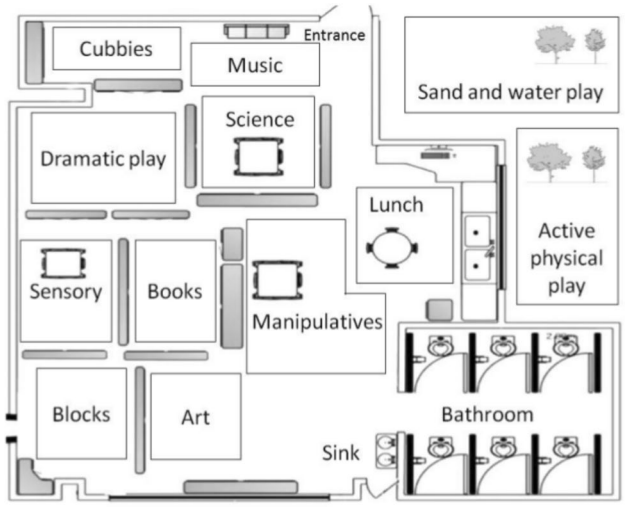

Fig. 6 Activity areas within a child care center

6 Location tracking system

In this section our aim is explore a novel childs location tracking approach, which can quantify childs vocal interaction in different activity/learning environments. A case study is presented from 5 children which include 3 males and 2 females. One of these children has a developmental delay while the remaining are typically developing. As shown in Table 5 two-fifth of these children have been exposed to non-English primary language. The data was collected in a high-quality child care center in the United States. As illustrated in Fig. 6, the center has 7 different activity areas: art, block, book, dramatic play, cubbies, manipulatives, and science.

After applying the child–adult turn-taking detection to the data recorded from the case study children, we managed to estimate the percentage of the time each child engages in the communication with other children and adults (Sect. 5.2). For each child, three location recordings

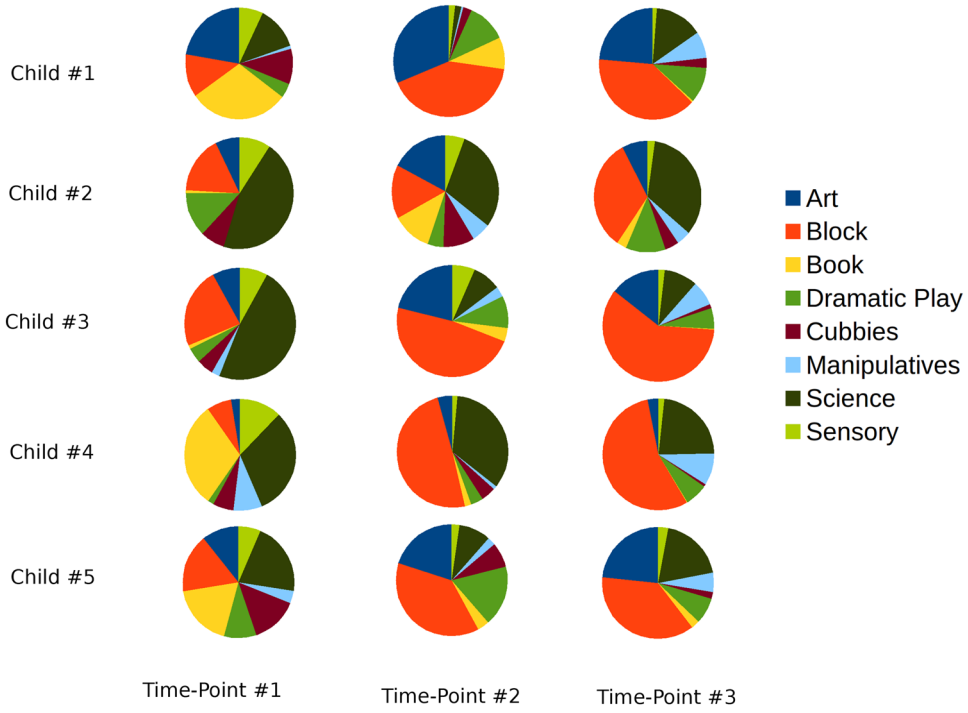

Fig. 7 Time (%) that five children spent at seven different activity areas at three different Time-Points

that can be captured in the i-Vector space. However, no dramatic change in error rate has been resulted after increasing the length of audio segments from 1.5 to 3 s. This may be due to the fact that, using longer audio segments (e.g. 2 and 3 s) will result in estimation of more informative i-Vector features (T-matrix with higher rank are found more effective; (Table 4). Hence, the improvements resulted by using i-Vectors estimated from longer segments has helped with compensating for the errors occurred as a result of missing the rapid turn-takings during the conversation (Table 4).

were chosen from three typical days at three different Time-Points within one classroom at a childcare center (giving us a total of 9 h of evaluation data per child). There were a number of none-speech occurrences during the recorded audio files (music, and crowd noise) in this study all those occurrences were removed. For Time-Point #1 and #3, audio files were recorded during the morning where the posted class schedule consisted of science, arts, blocks, and free play. For Time-Point #2 audio files were recorded during the afternoon where the schedule consisted of free play, art,

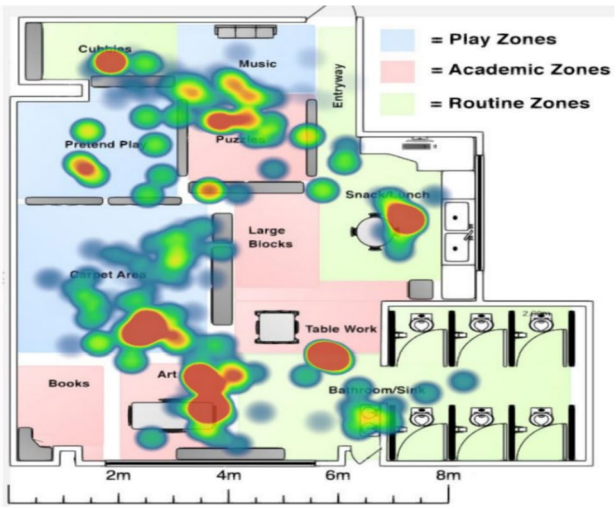

Fig. 8 Heat map adult word count vocalizations per minute

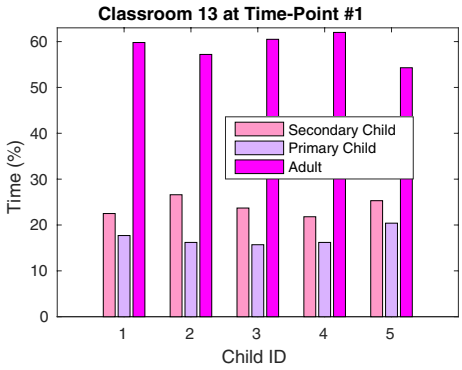

Fig. 9 Level of interaction between children and other children and adults, Time-Point #1

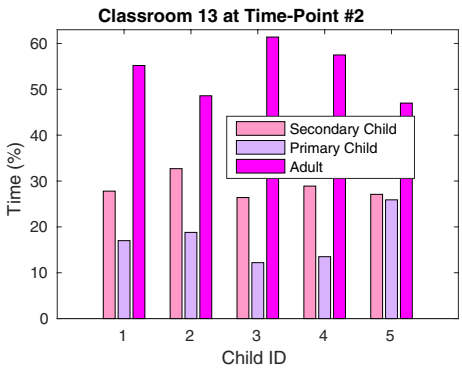

Fig. 10 Level of interaction between children and other children and adults, Time-Point #2

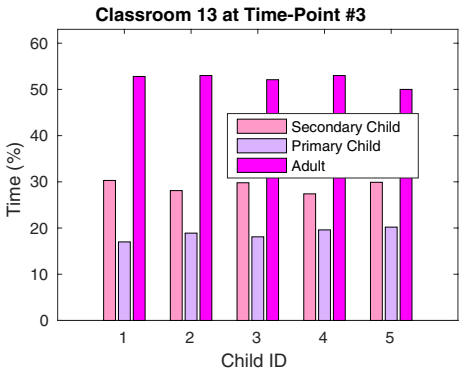

Fig. 11 Level of interaction between children and other children and adults, Time-Point #3

hand washing, having snacks, nap and quite time. Analysis are carried out during three distinct Time-Points. The 3 h Time-Points shown in Figs. 7, 8, 9, 10, and 11 are not necessarily synced. This is because the non-speech frames from the audio labels and also missing locations from the location labels are removed from the analysis. The experiments take into account:

- Primary child: speech initiated by the child wearing the LENA unit.
- Secondary child: speech originated by other children and directed at the primary child within his/her close proximity.
- Adult: speech originated by an adult and directed at the primary child within his/her close proximity.

In the following sections the effect of different language environment for verbal communication quantity is explored. We present an analysis of children vocal interaction quantity through location tracing, that is based on three types of observations:

- Time that children spent at different learning areas (Sect. 6.1).
- Adult-child verbal interaction level with respect to different learning environments (Sect. 6.2).
- Quantity of verbal communication (i) child–child, (2) child–adult (Sect. 6.3).

Finally, analysis and discussions taking into account all these observations are provided in Sect. 6.4.

6.1 Children's time spent in language environments

In this section we analyze the child's language environment, and study at what percentage of time each child is vocalizing in each activity area. Understanding which activity areas

stimulate verbal communication can be helpful for more efficient language development.

Figure 7 illustrates time, which children spent at different activity areas, at different Time-Points. The statistics are collected from 5 children, during Time-Points #1, #2 and #3. It is interesting to make the following observations:

(1) Time spent by each child in different learning areas during different Time-Points. It can be seen that Child #1, Child #4, Child #5 during Time-Point #1 spent a fair amount of time in the area of books, but during Time-Points #2, #3 in the area of block. Child #2 during all the Time-Points prefered the area of science. Analyzing the case of Child #3, it is seen that in the Time-Point #1 dominating zone was science, and during the other Time-Points it was area of block. Children spent least amount of their time in sensory and manipulatives activity areas.
(2) What learning areas are the most popular among children during different Time-Points. As it is illustrated in the Fig. 7, during the Time-Point #1 the most dominative are learning environments of books and science. All the children spent most of their time in the block area during Time-Point #2 and Time-Point #3.

6.2 Adult's vocal interaction level in different activity areas

An intervention is important to stimulate child's voice communication. In order to understand the need of child assistance, we explore in which learning areas children demand more verbal assistance. Figure 8 displays the heat map of the adult word count vocalizations per minute in each activity area. The hot red spots indicate the highest adult-to-child interaction lever, and the blue spots shows the lowest one. It can be observed that teachers were mostly vocalizing during the art activities that belongs to the academic zone. As well, teachers had hight vocal interactions in play zone, more specifically in books, science, and routine zone near the sink, lunch table, and cubbies area. In contrast to it, the teachers had the lowest verbal interaction level in blocks and manipulatives areas that belong to the academic zone.

6.3 Children's vocal interaction quantity during different time points

Figures 9, 10, and 11, help to understand the childs language environment by estimating the time of vocal interaction during Time-Point #1, Time-Point #2, Time-Point #3, respectively. We explore how much time in percentage each of five children spent talking. At the same time we take into account verbal communication of (i) secondary child, (ii) primary child, (iii) adult. The statistics shows the quantity of communication between children and with adults, that can help to explore which child need more teacher assistance. It can be observed that teachers spent more time interacting with child #4 during Time-Point #1, with child #3 during Time-Point #2. Meanwhile during Time-Point #3 all children interact with adults on average the same amount of time. In contrast to it, teachers direct the least amount of communication for child #5 during all Time-Points.

6.4 Analysis and discussions

In the previous subsections three types of observations to quantify the children vocal interaction were presented: (1) time that children spent at different learning environments, (2) quantity of vocal communication directed by adult in learning environments, (3) level of verbal communication between children, and children's communication with adults. In this section we relate and analyze all these different types of observations.

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. The illustrations show that during Time-Point #1 (Fig. 9), the average duration of conversation directed by the teacher to the primary child reaches its maximum and the average duration of conversation directed by other children to the primary child reaches a minimum compared to Time-Points #2 (Fig. 10) and #3 (Fig. 11). The amount of conversation directed to a child by the adult likely depend on the type of activity the child engages in. At Time-Point #1 the average duration of child–child interaction (conversation between primary and secondary children) is on average only 3.5 times higher than the amount of child–adult interaction (conversation directed at the child by the teacher).

This amount shows the nature of activity environment that child is exposed to was more teacher oriented compared to that of activities shown in Figs. 9, 10, and 11 for which this value reaches to 4.20 and 3.8 respectively.

During Time-Point #1 (Fig. 9), on average a smaller amount of speech was directed to child #4 by other children is bellow average. This may be because the childs primary language is not English and more efforts are needed to help engage him or her in conversations.

This can also be explained using information from Figs. 7, 8. Figure 7 shows that child #4 spent 31% of his time in the books and science areas respectively, while the other children spent a majority of their time in the blocks and science area. As it is illustrated in Fig. 8, the books activity area involves the higher amounts of speech from the teacher to the children. This may explain the smaller amount of child–child conversation for child #4 is due to the nature of activities he has been engaged to. Despite

the fact that child #2 had symptoms of development delays on average, a relatively high amount of conversation was between this child and other children and teachers.

During Time-Point #2 (Fig. 10) which included the afternoon activities, the amount of speech directed by teachers to children is reduced compared to Time-Point #1. This can be explained by Figs. 7 and 8. As it is shown in Fig. 7, the average amount of time spent by children in the science area was reduced considerably and now a majority of children spent up to 40 percent of their time in the block activity area. In 8 it is observed that in block area teachers almost do not vocaly interact with children.

Across all three Time-Points, the highest relative average amount of speech produced by primary and secondary children takes place during Time-Point #3 (Fig. 11). This may be explained by Fig. 7 that shows on average highest amount of time was spent on block activities (child–child communication oriented). Also during Time-Point #3 the average amount time spent in the book area is the lowest across all three time points, which may explain why the average amount of speech directed to children by adults is reduced compared to other Time-Points across all children.

We presented a novel approach, that helps to quantify a level of children vocal interaction through location tracking. Collectively, children's time spent in different areas (Fig. 7), the heat map of teachers vocal interaction (Fig. 8), and the Time-Points corresponding to 5 children in Figs. 9 10, and 11 allow us to gain a wider perspective of child communication with teachers and peers in the classroom across different activity areas. Our analysis plots support our ability to:

- Determine which children are less engaged in voice communication.
- Assess how much communication children have with other children in specific activity areas.
- Determine which activities stimulate greater voice communication between child–teacher and child–child.
- Determine which activity areas individual children or all children within a given classroom on average spent their time (e.g., on average the largest and smallest amount of time was spent in the blocks and manipulatives areas, respectively).

Providing teachers the information about the language environment, children experience and the locations they occupy, will allow early educators to better arrange interactions in the classroom that support childrens social and pre-academic learning. Our system provides a framework that may be useful in finding patterns in the global and local cues which provide the discrimination on the level of childs engagement in vocal communication.

7 Conlusion

In this paper we presented a language monitoring system using (1) a child–adult turn-taking, (2) a novel location tracking approach. A close to real-time (1.5 seconds delay) child–adult speech turn-taking system was introduced. Our experimental results show that applying TO-COMBO-SAD prior to i-Vector based classification results in up to 27.3% and 21.34% relative error rate reduction compared to the baseline results produced by the LIUM speaker diarization system and the baseline i-Vector based system respectively.

Looking at the confusion matrix of the TO-COMBO-SAD i-Vector based system it can be observed, 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. Comparing the confusion matrix of the TO-COMBO-SAD i-Vector based system with that of the baseline i-Vector system, it was showed that applying this speech activity detector is quite beneficial and results in reducing the number of errors occurred due to the confusion between speech and non-speech segments. It was also showed the effect of segment duration on classification performance. The best results were achieved by using segments with 1.5s duration. Using segments of shorter length limits the amount of useful speaker dependent information captured in the i-Vector space, and using longer segments will result in missing rapid turn-taking points.

A novel case study was presented to show the importance of speech and location analysis in building a foundation for the analysis of child language environment. We relied on three types of observations: (1) children's time spent in different language environments, (2) the heat map of adult's vocal interaction level in different activity areas, (3) children's vocal interaction quantity during different time points. These results can help us to understand where a child spends more/less time during specific classroom activities as well as the amount of verbalizations, both spoken and heard, which may affect their interest in specific activity areas.

Acknowledgements Authors wish to express our sincere thanks to Univ. of Kentucky for the joint collaboration efforts on this study. In particular, wish to thank Ying Luo for collecting, organizing the child database used in this study.

References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356–370.

Bahari, M. H., McLaren, M., van Leeuwen, D. A., et al. (2014). Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, *34*, 99–108.

Barras, C., Zhu, X., Meignier, S., & Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1505–1512.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: *Proceedings of the institute of phonetic sciences* (Vol. 17, pp. 97–110). Amsterdam.

Bonastre, J.-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N.W., Fauve, B.G., & Mason, J.S. (2008). ALIZE/spkdet: A state-of-the-art open source software for speaker recognition. In: Odyssey. p. 20.

Connaghan, D., Hughes, S., May, G., Kelly, P., Conaire, C.Ó., O'Connor, N.E., O'Gorman, D., Smeaton, A.F., & Moyna, N. (2009). A sensing platform for physiological and contextual feedback to tennis athletes. In: *Wearable and implantable body sensor networks, 2009* (pp. 224–229). BSN 2009. IEEE.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011a). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., & Dehak, R. (2011b). Language recognition via i-vectors and dimensionality reduction. In *Twelfth Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Delano, M., & Snell, M. E. (2006). The effects of social stories on the social engagement of children with autism. *Journal of Positive Behavior Interventions*, *8*(1), 29–42.

Gauvain, J.-L., & Lee, C.-H. (1991). Bayesian learning of Gaussian mixture densities for hidden Markov models. In *Speech and natural language: Proceedings of a Workshop Held at Pacific Grove, California, 19-22 February, 1991*.

Ghaemmaghami, H., Dean, D., & Sridharan, S. (2015). A cluster-voting approach for speaker diarization and linking of Australian broadcast news recordings. In *ICASSP* (pp. 4829–4833). IEEE.

Ghaemmaghami, H., Dean, D., Vogt, R. & Sridharan, S. (2011). Extending the task of diarization to speaker attribution. In *Interspeech 2011*, 28–31 August 2011, Florence.

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

Gravier, G., Betser, M., & Ben, M. (2010). AudioSeg: Audio segmentation toolkit, release 1.2. IRISA, January.

Gupta, R., Bone, D., Lee, S., & Narayanan, S. (2016). Analysis of engagement behavior in children during dyadic interactions using prosodic cues. *Computer Speech & Language*, *37*, 47–66.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.

Huijbregts, M. A.H. (2008). Segmentation, diarization and speech transcription: Surprise data unraveled. Ph.D. thesis, Centre for Telematics and Information Technology University of Twente.

Kasari, C., Gulsrud, A. C., Wong, C., Kwon, S., & Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. *Journal of Autism and Developmental Disorders*, *40*(9), 1045–1056.

Meignier, S., & Merlin, T. (2010). Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop* (Vol. 2010). Le Mans: Universite du Maine.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, *20*(2), 303–330.

Najafian, M., Irvin, D., Luo, Y., Rous, B.S., & Hansen, J.H. (2016). Employing speech and location information for automatic assessment of child language environments. In *Sensing, processing and learning for intelligent machines (SPLINE)*. IEEE, pp. 1–5.

Phebey, T. (2010). The Ubisense assembly control solution for BMW solution for BMW. *Proccedings of RFID Journal Europe Live*. Retrieved 18 August, 2016.

Reynolds, D.A., Singer, E., Carlson, B.A., O'Leary, G.C., McLaughlin, J.J., & Zissman, M.A. (1998). Blind clustering of speech utterances based on speaker and language characteristics. In *Fifth International Conference on spoken language processing—ICSP*.

Riehle, T.H., Lichter, P., Giudice, N.A. (2008). An indoor navigation system to support the visually impaired. In *Engineering in Medicine and Biology Society, 2008*. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, pp. 4435–4438.

Sadjadi, S. O., & Hansen, J. H. (2013). Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing Letters*, *20*(3), 197–200.

Safavi, S., Russell, M., & Jančovič, P. (2014). Identification of age-group from children's speech by computers and humans. In *Fifteenth Annual Conference of the International Speech Communication Association—INTERSPEECH*.

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *IEEE International Conference on acoustics, speech, and signal processing, 1997*. IEEE. ICASSP-97 (Vol. 2, pp. 1331–1334).

Siegler, M.A., Jain, U., Raj, B., & Stern, R.M., (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of DARPA speech recognition workshop*. Vol. 1997.

Swedberg, C. (2011). Bmw finds the right tool. *RFID Journal*, *1*, 2009.

Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1557–1565.

Vijayasenan, D., & Valente, F. (2012). Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In *Thirteenth Annual Conference of the International Speech Communication Association—INTERSPEECH*. Portland.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development*, *65*, 606–621.

Woźniak, M., Odziemczyk, W., & Nagórski, K. (2013). Investigation of practical and theoretical accuracy of wireless indoor positioning system ubisense. *Reports on Geodesy and Geoinformatics*, *95*(1), 36–48.

Yella, S. H. (2015). Speaker diarization of spontaneous meeting room conversations. PhD thesis, EPFL, Lausanne.

Zhao, Q., Kawamata, M., & Higuchi, T. (1988). Controllability, observability and model reduction of separable denominator MD systems. *IEICE Transactions (1976–1990)*, *71*(5), 505–513.

Ziaei, A., Kaushik, L., Sangwan, A., Hansen, J.H., & Oard, D.W. (2014). Speech activity detection for nasa apollo space missions: Challenges and solutions. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Ziaei, A., Sangwan, A., & Hansen, J.H. (2013). Prof-Life-Log: Personal interaction analysis for naturalistic audio streams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7770–7774). IEEE.