Audio-Based Group Detection for Classroom Dynamics Analysis

Yudong Tao*, Samantha G. Mitsven†, Lynn K. Perry†, Daniel S. Messinger† and Mei-Ling Shyu*
**Department of Electrical and Computer Engineering*
University of Miami, Coral Gables, USA
Email: {yxt128, shyu}@miami.edu
†Department of Psychology
University of Miami, Coral Gables, USA
Email: {sxm1531, lkperry, dmessinger}@miami.edu

***Abstract*—Group detection is a fundamental problem in sociological and behavioral data analysis and has attracted much attention in recent years. Most of the current approaches focus on using visual data, such as still images and videos, to detect groups. One of the most important applications of group detection is to assist psychologists to understand the classroom dynamics. However, the camera recordings may be unavailable and it could be infeasible to set up the cameras without blind spots. Therefore, as an alternative approach to group detection, we propose an audio-based framework that utilizes multiple synchronized audio data streams collected from wearable devices on each subject. In this paper, the audio recordings collected from a preschool classroom over multiple days are used to produce the group detection results which are validated by clustering the subject locations collected along with the audio data. The experiment shows on average 0.391 Normalized Mutual Information (NMI) scores for the detected groups by the audio-based framework and location-based clustering.**

***Keywords*-group detection, classroom dynamics analysis, audio processing**

I. Introduction

Detecting conversational groups has recently drawn much attention since it plays an important role in social group analysis, social robotics, and video surveillance [1]. Such results can be applied to a variety of application domains, such as improving human-robot interaction [2] and automated comprehension of social communications [3]. Moreover, the automated group detection results can provide objective and quantitative measurements for interactive human behavior analysis [4]. This particular analysis can be an initial step in understanding the classroom dynamics and the effects of social interactions on children's cognitive and social development.

Young children's interactions, both with teachers and peers, in early education programs have a long-lasting impact on their cognitive and social development. The exposure to the variegated and sophisticated vocabulary from teachers is related to the preschoolers' language gains, including the growth in syntactic comprehension [5] and oral language skills [6], as well as their later reading comprehension abilities [7]. Further, vocal turn-taking between teachers and children is related both to children's vocalizations in the moment and their vocabulary growth over the course of the school year [8]. Social interactions with peers provide children with the opportunities to learn through modeling and imitation [9] as well as to expand their social and cognitive knowledge through collaborations [10]. Interactive plays with peers in the classroom facilitate positive developmental outcomes in children's language [11], [12], social [13], learning [14], and cognitive competencies [15]. Thus, quantifying the simultaneous interactive dynamics between children and their teachers and peers within the preschool classroom is fundamental to understanding the individual developmental trajectories and can help elucidate the characteristics of interactions that can be leveraged to promote the optimal development in all children.

Understanding the relation between early social dynamics and later academic achievement is of particular importance for children exhibiting delays in development as it could provide an insight into features of interactions that would benefit from targeted intervention. Children with hearing loss experience an initial period of auditory deprivation where their access to spoken language is limited, potentially leading to delays in language use and acquisition. The participation in the oral language education programs positively impacts the oral communication abilities of children with hearing loss [16]. While previous research has suggested that children with hearing loss benefit from the exposure to the enriched language environment of the oral language education programs. What is missing from these accounts is a rigorous examination of the in-the-moment features of classroom dynamics within these intervention classrooms.

Since the spatial-temporal nature of the data increases its complexity, machine learning techniques are commonly applied to learn the pattern within the data [17]. Most of the current research focuses on free-standing conversational group detection (called F-formation detection) based on the visual data such as still images and videos. These approaches aim to localize o-space [18] which refers to a common space that all the subjects in the group have direct, equal, and exclusive access. Hung and Kröse first solved the problem with dominant sets [19]. Following their work, game theory [20], Hough voting [21], and graph clustering [1] algorithms were proposed to improve the performance of F-formation group detection. Ketti et al. [22] presented a graph-cut algorithm that considers both probabilities of groups and the visibility constraints, and achieved the currently best among all the existing techniques. Recently, deep learning has also been exploited for F-formation group detection [23].

While visual data has been exploited to detect interactions, its use has certain limitations. First, to record a room without blind spots, multiple cameras are needed and those cameras need to be configured correctly. It could also be infeasible to setup cameras in the environment, for example, when there are open areas without fixtures to place the cameras to or when the privacy is of concern. Meanwhile, visual data can only provide information whether people gather in space. It implies that the subjects share a common space for conversation in the context of cocktail banquet or poster section. However, in a general environment, such as classroom, individuals could stand together without intention to communicate. A more straightforward approach is to use audio recordings to identify instances where the subjects are involved in the same conversation. Therefore, in this paper, we propose a group detection framework based on the audio data collected by the audio recorders worn by the individual subjects within a classroom to identify those who share the same vocal input and those who have conversations with each other.

The rest of this paper is organized as follows. Section II introduces the related work and techniques of group detection and classroom dynamics analysis. Our proposed audio-based group detection framework for classroom dynamics analysis is presented in Section III. Section IV illustrates the empirical experimental results for multi-day classroom data. Finally, Section V highlights our conclusions and discusses the future directions.

II. Related Works

A. Group Detection

Group detection based on visual data has been actively studied as the F-formation detection problem [18]. The dominant set approach was first developed in 2011 and considers the affinities between subjects in the images as the weights of a graph [19]. This idea is widely used to study the correlation among semantic concepts [24] and was also used in the later research while the method to determine the affinity between subjects and the method to form the groups are improved.

The game-theory for conversational group approach proposed by Vascon et al. [20] utilizes a statistical model to compute the affinity based on the possibility of two subjects sharing attention and the evolutionary game theory to consider the temporal information to refine the affinity. Papers [1] and [21] utilized the graph clustering techniques and Hough voting strategy respectively to determine which subjects belong to the same group. Ketti et al. [22] proposed the graph-cut algorithm to integrate positions, orientations, and visibility constraints to determine the groups, which can achieve the state-of-the-art performance.

Other than the visual-based approaches, group detection is also exploited based on the location sensors [25], which focuses on the subjects sharing the same trajectory over time. In this approach, data collected from multiple sensors are leveraged to identify the trajectory of individual subjects using the Kalman filter and detect the groups with relative clustering techniques.

B. Classroom Dynamics Analysis

Much of the previous work that studied classroom dynamics among children relied on relied on teachers' observations, interviews with children, as well as experts' manual coding of classroom interactions. Both teacher ratings of the frequency with which children play or have conflict with one another as well as the students' own reports of peers who they do and do not like to play with have been used to construct classroom networks of interactions [26], [27]. While teacher observations and children's self-reported friendships can provide global information regarding the peers who the children generally play or have conflict with, they miss out on the moment-to-moment or day-to-day variation in interactions and friendships.

Manual coding schemes, on the other hand, involve observing one child at a time and recording the features of their individual interactions, including the valence of their interactions (positive vs. negative) as well as their proximity to each of their peers [27]. While these coding schemes provide rich information for individual children over short intervals of time (i.e., 25 minutes of the observation), they are unlikely to capture all the interactions going on at the moment. Researchers have also employed head-mounted cameras worn by children in the classroom to record their linguistic environments from the first person perspective, i.e., individuals' language inputs and uses [28]. However, this analyzing data collected from this measurement technique is labor intensive, requiring manual transcriptions of multi-hour recordings.

To overcome the labor intensive nature of manual coding of interactions and to broaden the scope of simultaneous, moment-to-moment information that can be acquired, researchers have begun to apply automated data collection techniques to understand the classroom dynamics. The LENA system, which is composed of lightweight audio recorders and pattern recognition software, allows for the continuous, automated collection and analysis of children's language experiences, including both their language input from peers and adults and their own language use, in everyday contexts like the classroom [8], [29]. Radio frequency identification (RFID) technology has also been employed in the classroom to capture continuous measurements of children's location and movement to generate the measures

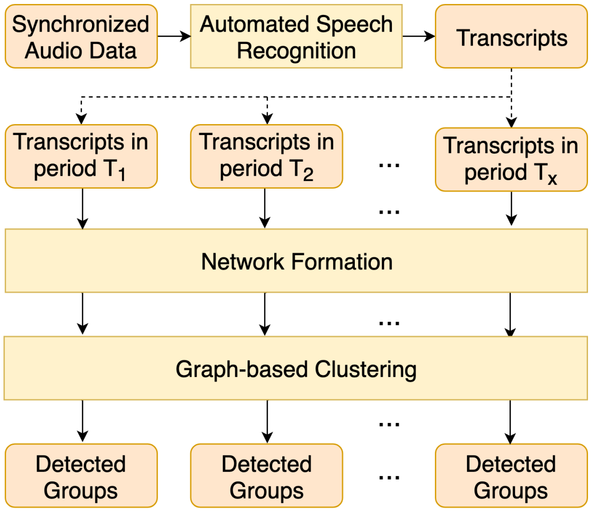


Figure 1. The proposed audio-based group detection framework

of the velocity of movements, and to indicate when children are in social contact [30].

III. Audio-based Group Detection Framework

A. Overall Framework

In this paper, we focus on the proposed group detection framework based on audio data. As shown in Figure 1, the proposed audio-based group detection framework includes three steps, automated speech recognition (ASR), network formation based on the transcripts, and graph-based clustering. The inputs of this framework are a set of audio files which are collected by wearable devices on each subject in the room, and these audio files are synchronized in time, i.e., all the audios start and end at the same time. After that, each audio recording is transcribed by the ASR engine and converted to texts with timestamps, indicating the beginning and end of the transcript. The text data of individual subjects are then analyzed to compute the affinity among the subjects and form the network, where the connection between subjects are represented by the weights of the graph. In the end, the graph-based clustering algorithm is used to produce the groups for each time period. In particular, the self-tuning spectral clustering algorithm [31] is applied to automatically determine the number of clusters (i.e., number of groups). In this paper, the groups will be identified at every observation period of t seconds and the gap between two adjacent observation periods is denoted as Δt seconds. Therefore, the i-th observation periods can be defined as $T_i = [(i-1)\times\Delta t, (i-1)\times\Delta t + t], i = 1, 2, \ldots$. All the subjects will be assigned to one and only one group for each observation period and all the subjects in the same group are supposed to share a common conversational environment in the specific period of time.

B. Automated Speech Recognition

Based on our definition of a group, the subjects in the same group should be involved in the same conversation. Therefore, in the proposed framework, audio data are first converted to transcripts with the pre-trained ASR model. Since all the audio data collected by the wearable devices could include various types of noise and could be affected by many unmanageable factors, this step could also benefit the overall performance by mitigating the effect of the noise in the audio recordings. We assume that the subjects are in the same conversational environment if their recorders contain the same conversation, i.e., have the same or similar transcripts.

Automated speech recognition has been studied in the past decades. Due to the recent advances in deep learning [32], the performance of the ASR system has been improved significantly and is ready to provide stable results [33]. Deep learning models are able to learn auditory features from large amounts of data and thus illustrate superior performance in all benchmarks [34]. In this study, we apply Google Speech-to-Text cloud service[1] to produce the audio transcription results. The output of the ASR module include the transcripts and the start time and end time of each word in the transcripts. Therefore, we denote the outputs of the ASR system as a list of words $w_1, w_2, \ldots, w_N$ and their corresponding timestamps $(t_{s,1}, t_{e,1}), (t_{s,2}, t_{e,2}), \ldots, (t_{s,N}, t_{e,N})$, where w_k is the k-th word in the transcript, N is the number of words in the transcript, and $t_{s,k}$ and $t_{e,k}$ are the start time and end time of w_k. Meanwhile, the transcription ensures that there is no overlapping time period of any two words, i.e., $\forall k_1, k_2 \in [1, N], k_1 \neq k_2, (t_{s,k_1}, t_{e,k_1}) \bigcap (t_{s,k_2}, t_{e,k_2}) = \emptyset$. In other words, the words in the transcript come successively, i.e., $\forall k_1 < k_2 \in [1, N], t_{s,k_1} < t_{e,k_1} < t_{s,k_2} < t_{e,k_2}$.

C. Transcript-based Network Formation

Among all the subjects in the experiments, the relation between each pair of subjects can be regarded as an edge in the graph while each vertex represents a subject. Therefore, the relationship among all subjects can be naturally represented by a weighted graph $G = (V, E, W)$, where V refers to the set of vertices, E refers to the set of edges, and W refers the weights of the edges in graph G. The weights of each edge represents the affinity of a pair of subjects connected by the edge. Without loss of generality, we assume that the higher the affinity between the subjects, the smaller weight is assigned. In Figure 2, an example network of six subjects is depicted. The numbers on the edges indicate the affinity between two connected subjects. Subjects C and E have the largest weight and thus they have the lowest affinity.

In order to compute the affinity between subjects in every observation period, the transcripts involved in the period is

[1] https://cloud.google.com/speech-to-text/

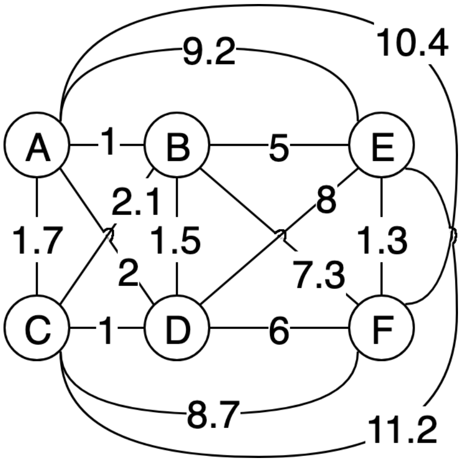


Figure 2. An example of the network representation

utilized as the inputs for the analysis. For the i-th observation period, the words outside of T_i are first filtered out. The word w_k will be included if it satisfies the constraint below.

$$(t_{s,k}, t_{e,k}) \bigcap T_i \neq \emptyset. \tag{1}$$

We denote the list of words from the transcripts of subject j as L_j. Thus, the affinity between subject j_1 and j_2 can be represented by the distance between L_{j_1} and L_{j_2}, denoted as $d(L_{j_1}, L_{j_2})$. To appropriately measure the distance, we propose to use the Manhattan distance between the word count vectors,

$$d_M(L_{j_1}, L_{j_2}) = \sum_{w \in \mathbb{W}} (|N_{j_1}(w) - N_{j_2}(w)|), \tag{2}$$

where $\mathbb{W}$ is the set of words presented in either L_{j_1} or L_{j_2}, and $N_{j_1}(w)$ and $N_{j_2}(w)$ are the word count of w in L_{j_1} and L_{j_2}, respectively.

Alternatively, Levenshtein distance between two lists of words [35] can be used to evaluate how similar two transcripts are. The Levenshtein distance can be computed by $d_L(L_{j_1}, L_{j_2}) = d_{\text{lev}}(p, q)\Big|_{p=|L_{j_1}|, q=|L_{j_2}|}$, where

$$d_{\text{lev}}(p,q) = \begin{cases} \max(p,q) & \text{if } \min(p,q) = 0 \\ \min \begin{cases} d_{\text{lev}}(p-1,q)+1, \\ d_{\text{lev}}(p,q-1)+1, \\ d_{\text{lev}}(p-1,q-1)+I_{p,q} \end{cases} & \text{otherwise} \end{cases}$$

$|L_j|$ is the number of words in L_j, $I_{p,q}$ is 1 if the p-th word in L_{j_1} and the q-th word in L_{j_2} are different and $I_{p,q}$ is 0 otherwise. Here, p and q are non-negative integers ranged in $[0, |L_{j_1}|]$ and $[0, |L_{j_2}|]$, respectively. Levenshtein distance is commonly used in approximate string matching and the evaluation of automated speech recognition, and it measures the minimal number of edits (insertion, deletion, and substitution) of words required to make L_{j_1} and L_{j_2} the same.

Algorithm: STSC for group detection

```
Inputs: Affinity matrix $A \in \mathbb{R}^{S \times S}$
Outputs: Number of groups $N_g$ and the subjects in each group
1: Compute the normalized affinity matrix
   $\hat{A} = D^{-1/2} A D^{-1/2}$
2: Compute $X = [x_1, x_2, \ldots, x_S]$, where $x_1, \ldots, x_S$ are the eigenvectors of $\hat{A}$
3: for $c = 1$ to $S$ do
4:     Calculate $R_c = \arg\min_{R_c} \text{loss}_c$ by the gradient decent algorithm, where $\text{loss}_c$ is defined in Eq. (3)
5:     Compute the minimized cost $\text{loss}_c$ based on Eq. (3), given the optimized rotation $R_c$
6: end for
7: Determine the optimal number of groups by
   $N_g = \arg\min_c \text{loss}_c$
8: Compute $Z^* = X R_{N_g}$
9: for $i = 1$ to $S$ do
10:    Assign subject $i$ to the $k_i$-th cluster, where
       $k_i = \arg\max_j {Z^*_{ij}}^2$
11: end for
```

Figure 3. Description of STSC Group detection algorithm

Although d_M is simpler to compute and compare the similarity of each transcript pair, it neglects the order of the words in the transcripts. Therefore, d_L should better reflect the affinity between subjects.

D. Group Detection

Once the affinity is computed and the network graph is formed, a group detection can be performed using the clustering algorithm. In this study, we apply spectral clustering to detect the groups sharing the same conversational environment in each time period, which considers the network structure and minimizes the in-group Levenshtein distance between the subject pair. Since in the classroom dynamics analysis, it is almost impossible to manually determine the number of groups in each time period, it is essential to determine the number of clusters automatically. Therefore, instead of regular spectral clustering, Self-Tuning Spectral Clustering (STSC) [31] is able to estimate the optimal number of clusters based on the eigenvectors of the affinity matrix A, where $A = \{a_{j_1 j_2}\}^{S \times S}$, $a_{j_1 j_2} = d(L_{j_1}, L_{j_2})$ and S is the number of subjects in the experiment. The STSC algorithm for group detection is summarized in Figure 3, where D in step 1 is a diagonal matrix and the i-th element in the diagonal can be computed by $D_{ii} = \sum_{j=1}^{S} a_{ij}$. The recovery of rotation R_c as mentioned in step 4 of STSC algorithm is conducted by the gradient decent method, which minimizes the objective function:

$$\mathrm{loss}_c = \sum_{i=1}^{S} \sum_{j=1}^{c} \frac{Z_{ij}^2}{M_i^2}, \tag{3}$$

where $Z = XR_c$, X is the column-wise concatenated eigenvectors, and $M_i = \max_j Z_{ij}$.

For each time period, the proposed framework is able to generate the number of groups for that time period and the subjects in each group.

IV. Experiments and Evaluations

A. Data Collection

In order to validate the effectiveness of our proposed framework, the audio data collected in multiple days of a classroom are used in the experiment. All the data were collected from children (ages 2.5 to 3.5 years old) enrolled in an English-dominant oral language inclusion classroom for children with hearing loss and their teachers.

The audios from all children and teachers were recorded using LENA Digital Language Processors (DLPs). Continuous measurements of children's location were collected using the Ubisense Dimension4 system. The Ubisense system is composed of four sensors, one in each corner of the classroom, which are linked by a timing and network cable. The Ubisense sensors provide data to a dedicated laptop running the Ubisense software, and track active tags (radio impulses emitting) worn by the children and teachers. The radio signal emitted by each tag was used to locate children in three-dimensional space by means of triangulation (angle of arrival = AoA) and time differences in arrival (TDoA). The Ubisense system has the capability to track up to 40 participants 4 times per second and the accuracy of 15 cm in the three-dimensional space.

Children wore LENA audio recorders in specially designed vests that contained a pocket on the front to house the recording device. In addition to the LENA recorder, children also wore Ubisense tags in pockets, one on the left and right sides of the vest. Teachers wore the LENA recorders and Ubisense tags in fanny packs worn around their waist. Ubisense and LENA recordings were collected simultaneously once per week for the entirety of the school day (approximately 4 hours) for 10 consecutive weeks during the spring semester (March-May of 2017). On recording days, there was an average of eight children and three adults (one primary teacher and two aides) in attendance. Because individual attendance varied, children contributed an average of 8 recordings (SD = 1.69) to analyses.

B. Network in Classroom

As shown in Figure 4, the Levenshtein distance between each pair of subjects in the classroom in one example day is presented, where the x-axis shows the time period over the day ($t = 180, \Delta t = 60$) and the y-axis shows the Levenshtein distance between the paired transcripts in each time period. On that day, 10 children and all three teachers were present in the classroom. Each line represents the trend of Levenshtein distance for a specific pair of subjects. The pair of subjects is more likely to be in the same group when the Levenshtein distance is closer to zero. On the other hand, it can be clearly observed that the trend lines of some subject pairs are clustered together at similarly high Levenshtein distances, which potentially indicates that the subjects in two groups are having different conversations.

Table I
The comparison of NMI and AMI scores between group detection of baseline methods and spatial distance

Features	NMI	AMI
Difference in MFCC	0.218	0.097
Difference in Spectrogram	0.273	0.125
Manhattan distance of word count vector	0.388	0.179
Levenshtein distance of transcripts	0.391	0.191

C. Group Detection

In Figure 5, each marker represents the average location of a subject in the classroom in the time period. The marker with the same color belongs to the same group. Figure 5(a) shows that the audio-based group detection framework clusters the subjects that are close to one another together. Meanwhile, Figure 5(b) shows a failure case, where the audio recording of subject 1 is not correctly transcribed due to the mechanical noise in the audio and thus not being clustered with subjects 0, 2, and 4 correctly.

Furthermore, to validate the effectiveness of our proposed framework, the location information of each subject in the classroom collected from the Ubisense data is used as an input to the STSC algorithm to generate the clustering results and this cluster is used as the ground truth to compare the performance of our proposed framework with some baseline methods. The normalized mutual information (NMI) scores and the adjusted mutual information (AMI) scores [36] between the spatial-distance-based groups and the groups generated by other algorithms are compared. The methods to be compared include (1) Euclidean distance of the Mel Frequency Cepstral Coefficients (MFCC) features of the audio data between two subjects [37], (2) Euclidean distance of the Short-Time Fourier Transform (STFT) Spectrogram, (3) Manhattan distance of the word count vector, and (4) Levenshtein distance of the transcripts. The average NMI and AMI scores are shown in Table I and our proposed framework shows much higher mutual information scores compared to the baseline methods (the second and third rows of Table I).

V. Conclusion and Future Works

In this paper, an audio-based group detection framework is proposed for classroom dynamics analysis. The proposed

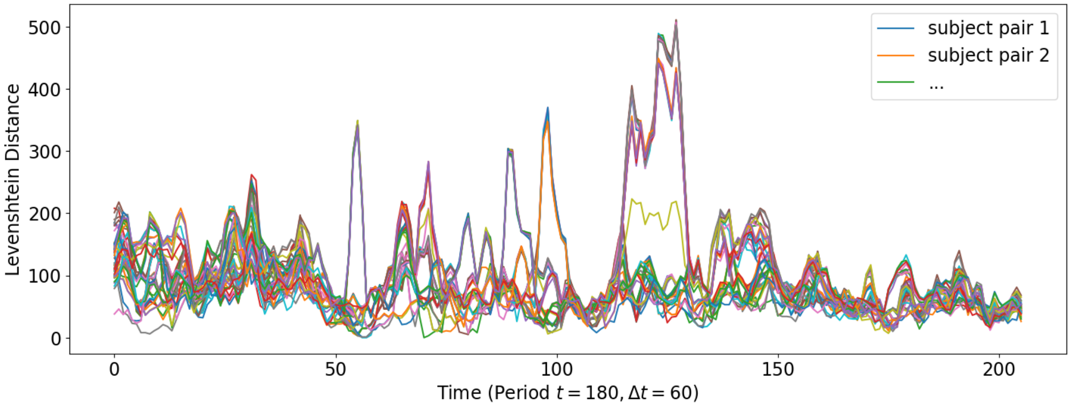
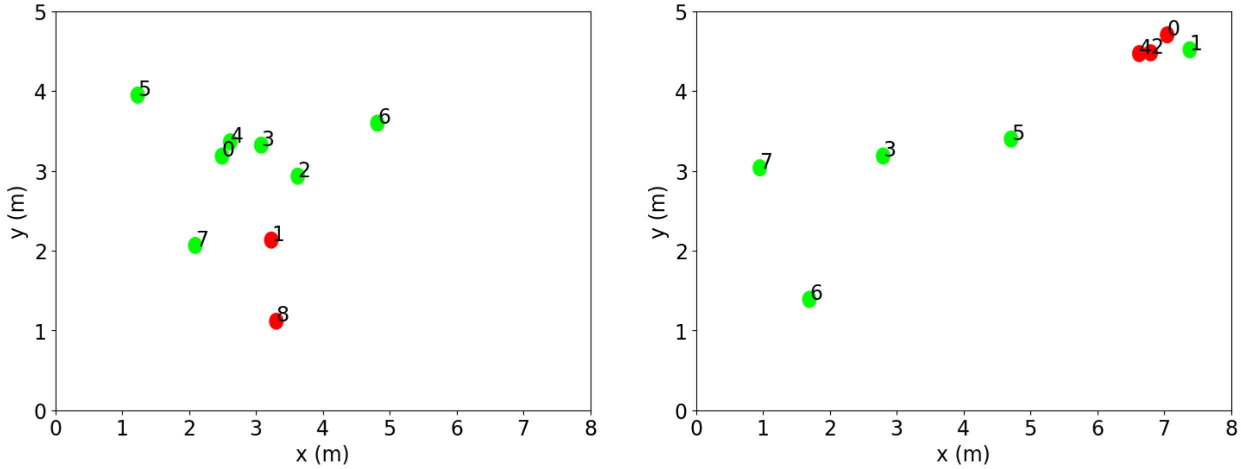


Figure 4. One-day Levenshtein distance trend for all subject pairs



(a) Success Case of group detection

(b) Failure Case of group detection

Figure 5. Visualization of example group detection results based on audio data in a specific time period

framework is able to automatically detect groups based on the audio data. These results could be used to provide quantitative measurements for classroom dynamics and help behavior analysis and classroom-based research. Based on our experimental results, our proposed framework can detect groups based on the audio data and outperforms the baseline methods when the Ubisense data is used as the reference.

In the future, the performance of our proposed framework can be further improved by incorporating different techniques. First, the temporal information can be considered. As discussed in Section II, game theory and other temporal smoothing methods can be applied to integrate the information over the time. Second, in addition to using the location data as the ground truth, the orientation of the subjects can also be considered. Third, the audio features such as MFCC can be integrated into the proposed framework to improve the robustness of audio-based group detection.

Acknowledgment

The research reported in this paper was supported by grants from the National Science Foundation (1620294) and the National Center for Special Education Research, Institute of Education Sciences Award (R324A180203).

References

[1] S. Inaba and Y. Aoki, “Conversational group detection based on social context using graph clustering algorithm,” in *International Conference on Signal-Image Technology & Internet-Based Systems*, Naples, Italy, November 2016, pp. 526–531.

[2] R. Triebel, K. O. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. A. I. Ramírez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. J. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang, “SPENCER: A socially aware service robot for passenger guidance and help in busy airports,” in *Field and Service Robotics - Results of the 10th International*

Conference. Toronto, Canada: Springer, June 2015, pp. 607–622.

[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transaction on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[4] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[5] J. Huttenlocher, M. Vasilyeva, E. Cymerman, and S. Levine, "Language input and child syntax," *Cognitive Psychology*, vol. 45, no. 3, pp. 337–374, 2002.

[6] P. B. Gámez, "Classroom-based english exposure and english language learners' expressive language skills," *Early Childhood Research Quarterly*, vol. 31, pp. 135–146, 2015.

[7] D. K. Dickinson and M. V. Porche, "Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities," *Child development*, vol. 82, no. 3, pp. 870–886, 2011.

[8] L. K. Perry, E. B. Prince, A. M. Valtierra, C. Rivero-Fernandez, M. A. Ullery, L. F. Katz, B. Laursen, and D. S. Messinger, "A year in words: The dynamics and consequences of language experiences in an intervention classroom," *PloS one*, vol. 13, no. 7, p. e0199893, 2018.

[9] A. Bandura, *Social learning theory*. New York: General Learning Press, 1971.

[10] L. S. Vygotsky, *Mind in society: The development of higher psychological processes*. Harvard university press, 1978.

[11] L. M. Justice, J. A. Logan, T.-J. Lin, and J. N. Kaderavek, "Peer effects in early childhood education: Testing the assumptions of special-education inclusion," *Psychological Science*, vol. 25, no. 9, pp. 1722–1729, 2014.

[12] Y. Rafferty, V. Piscitelli, and C. Boettcher, "The impact of inclusion on language development and social competerne among preschoolers with disabilities," *Exceptional Children*, vol. 69, no. 4, pp. 467–479, 2003.

[13] R. J. Bulotsky-Shearer, P. H. Manz, J. L. Mendez, C. M. McWayne, Y. Sekino, and J. W. Fantuzzo, "Peer play interactions and readiness to learn: A protective influence for african american preschool children from low-income households," *Child Development Perspectives*, vol. 6, no. 3, pp. 225–231, 2012.

[14] K. Coolahan, J. Fantuzzo, J. Mendez, and P. McDermott, "Preschool peer interactions and readiness to learn: Relationships between classroom peer play and learning behaviors and conduct." *Journal of Educational Psychology*, vol. 92, no. 3, pp. 458–465, 2000.

[15] G. B. Ramani, E. Zippert, S. Schweitzer, and S. Pan, "Preschool children's joint block building during a guided play activity," *Journal of Applied Developmental Psychology*, vol. 35, no. 4, pp. 326–336, 2014.

[16] J. S. Moog, "Changing expectations for children with cochlear implants," *Annals of Otology, Rhinology & Laryngology*, vol. 111, no. 5_suppl, pp. 138–142, 2002.

[17] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management*, vol. 6, no. 1, pp. 1–18, 2015.

[18] T. M. Ciolek and A. Kendon, "Environment and the spatial arrangement of conversational encounters," *Sociological Inquiry*, vol. 50, no. 3-4, pp. 237–271, 1980.

[19] H. Hung and B. Kröse, "Detecting f-formations as dominant sets," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 231–238.

[20] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "A game-theoretic probabilistic approach for detecting conversational groups," in *Asian conference on computer vision*. Springer, 2014, pp. 658–675.

[21] F. Setti, H. Hung, and M. Cristani, "Group detection in still images by f-formation modeling: A comparative study," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[22] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015.

[23] M. Swofford and J. Peruzzi, "Conversational group detection with deep convolutional networks," *CoRR*, vol. abs/1810.04039, 2018.

[24] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, "AAFA: Associative affinity factor analysis for bot detection and stance classification in twitter," in *IEEE International Conference on Information Reuse and Integration*. IEEE, 2017, pp. 356–365.

[25] S. Li, Z. Qin, and H. Song, "A temporal-spatial method for group detection, locating and tracking," *IEEE Access*, vol. 4, pp. 4484–4494, 2016.

[26] J. Chen, T.-J. Lin, L. Justice, and B. Sawyer, "The social networks of children with and without disabilities in early childhood special education classrooms," *Journal of autism and developmental disorders*, vol. 49, no. 7, pp. 2779–2794, July 2017.

[27] A. J. Santos, J. R. Daniel, C. Fernandes, and B. E. Vaughn, "Affiliative subgroups in preschool classrooms: Integrating constructs and methods from social ethology and sociometric traditions," *PloS one*, vol. 10, no. 7, p. e0130932, 2015.

[28] L. J. Chaparro-Moreno, L. M. Justice, J. A. Logan, K. M. Purtell, and T.-J. Lin, "The preschool classroom linguistic environment: Children's first-person experiences," *PloS one*, vol. 14, no. 8, p. e0220227, 2019.

[29] M. Soderstrom and K. Wittebolle, "When do caregivers talk? the influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments," *PloS one*, vol. 8, no. 11, p. e80646, 2013.

[30] D. S. Messinger, E. B. Prince, M. Zheng, K. Martin, S. G. Mitsven, S. Huang, T. Stölzel, N. Johnson, U. Rudolph, L. K. Perry *et al.*, "Continuous measurement of dynamic classroom social interactions," *International Journal of Behavioral Development*, vol. 43, no. 3, pp. 263–270, 2019.

[31] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, Canada, December 2004, pp. 1601–1608.

[32] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, p. 92, 2018.

[33] F. Dernoncourt, T. Bui, and W. Chang, "A framework for speech recognition benchmarking." in *Interspeech*, 2018, pp. 169–170.

[34] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, April 2018, pp. 4774–4778.

[35] M. J. Hunt, "Figures of merit for assessing connected-word recognisers," *Speech Communication*, vol. 9, no. 4, pp. 329–336, 1990.

[36] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Scientific reports*, vol. 6, p. 30750, 2016.

[37] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International Conference on Computer Graphics, Simulation and Modeling*, 2012, pp. 135–138.