A Speaker Diarization System for Studying Peer-Led Team Learning Groups

Harishchandra Dubey, Lakshmish Kaushik, Abhijeet Sangwan, John H. L. Hansen+

Center for Robust Speech Systems, Eric Jonsson School of Engineering
The University of Texas at Dallas, Richardson, TX 075080, USA
{harishchandra.dubey, lakshmish.kaushik, abhijeet.sangwan, john.hansen}@utdallas.edu

Abstract

Peer-led team learning (PLTL) is a model for teaching STEM courses where small student groups meet periodically to collaboratively discuss coursework. Automatic analysis of PLTL sessions would help education researchers to get insight into how learning outcomes are impacted by individual participation, group behavior, team dynamics, *etc.*. Towards this, speech and language technology can help, and speaker diarization technology will lay the foundation for analysis. In this study, a new corpus is established called CRSS-PLTL, that contains speech data from 5 PLTL teams over a semester (10 sessions per team with 5-to-8 participants in each team). In CRSS-PLTL, every participant wears a LENA device (portable audio recorder) that provides multiple audio recordings of the event. Our proposed solution is unsupervised and contains a new online speaker change detection algorithm, termed G^3 algorithm in conjunction with Hausdorff-distance based clustering to provide improved detection accuracy. Additionally, we also exploit cross channel information to refine our diarization hypothesis. The proposed system provides good improvements in diarization error rate (DER) over the baseline LIUM system. We also present higher level analysis such as the number of conversational turns taken in a session, and speaking-time duration (participation) for each speaker.

Index Terms: LENA, Naturalistic Audio Analysis, Speaker Diarization, Peer-led Team Learning (PLTL), Social Signal Processing.

1. Introduction

Peer-led team learning (PLTL) is a strategy used for improving learning outcomes in group settings for STEM students. Each team is led by a student who has already completed the course and is familiar with the course learning goals and/or challenges. The team lead coordinates the discussion on solutions of a given set of questions in study sessions. There is typically weekly study sessions held throughout the semester. PLTL is a popular approach and has been adopted by various universities at the undergraduate level. Additionally, education researchers have studied various aspects of PLTL to understand its impact on student's knowledge measured in terms of their success in academic programs [1, 2]. Typically, such research studies use control groups (by comparing students who do and do not participate in PLTL) and outcome metrics such as course grades or potentially opinions surveys to understand the educational impact. Here, analyzing actual student-to-student voice interaction in study sessions can help develop a richer understanding of how student success is related to participation, engagement,

group behavior, team lead, benefits *etc.*. However, this would require analyzing large quantities of data and the use of speech and language processing tools would be especially beneficial.

In this study, we explore the utility of speaker diarization technology in measuring simple communication metrics for PLTL sessions. Specifically, we describe a new corpus called CRSS-PLTL that was developed to facilitate this study. In CRSS-PLTL, we collected longitudinal data from 5 PLTL teams for one semester. Every PLTL session lasted for about 80 minutes where each team member wore a LENA audio recording device. Hence, the corpus contains multi-channel audio data for all sessions. This is different from typical diarization research that focuses on data collected using a single or multiple fixed far-field microphones [3, 4, 5]. It is common for students to physically move during PLTL sessions(*e.g.*, walking to whiteboard to solve problems) as well as breaking-up into smaller groups for discussion. The speaking style is spontaneous and casual. Short conversation turns and overlapped speech are often encountered. All these factors make speaker diarization challenging for these scenarios.

Speaker diarization systems have been extensively researched, often for specific tasks [6, 7, 8, 9]. Both supervised and unsupervised methods have been explored. Quiet recently, some researchers have suggested a method for speaker diarization using Restricted Boltzmann Machines [10]. The unsupervised methods for classification and segmentation of audio data has attracted attention in recent years [11]. Among multi-stream diarization, meeting recordings have been analyzed by combining MFCC and TDOA features with various segmentation and clustering algorithms [3, 4, 5]. In this study, we propose an unsupervised system for diarization suitable for studying PLTL groups. Particularly, we propose new unsupervised methods for speaker change detection and speaker clustering. In our experiments, we compared the proposed method with the LIUM diarization system. The proposed method in this study achieves more than 10% absolute reduction in diarization error rate (DER) over LIUM for CRSS-PLTL data. Finally, we also use the diarization information to compute downstream metrics such as the number of conversational-turns taken and participation, and discuss how such metrics can assist in automatic analysis of PLTL groups.

2. Proposed System

Fig. 1 shows the proposed system. As shown in the figure, each PLTL team member wears a LENA audio recorder unit (that essentially acts as a close-talk microphone). Therefore, each session yields multi-channel audio data where the number of channels was equal to the number of participants. This makes CRSS-PLTL corpus somewhat different from corpora typically used for diarization research where fixed far-field microphones are used for audio capture. This difference allows us to solve
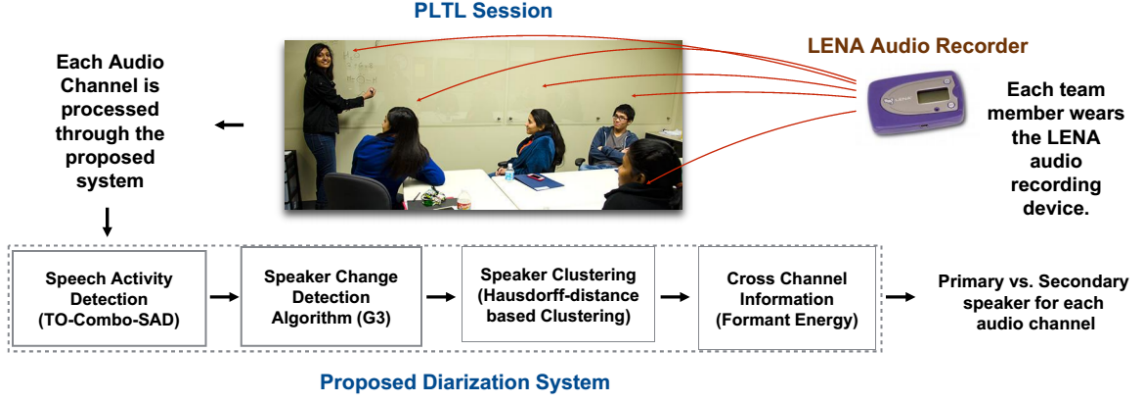
Figure 1: *In CRSS-PLTL, LENA audio recorders were worn by each team member for the entire session, that yields multi-channel audio data. The proposed speaker diarization system uses TO-Combo-SAD [12] to remove non-speech segments, and then uses Unsupervised G^3 algorithm along with Hausdorff-distance based clustering to perform speaker change detection and clustering, respectively.*

the overall diarization problem by solving primary speaker (person wearing the LENA device) *vs.* secondary speaker (all other speakers) detection problem for each audio stream. In other words, we were always solving a two-speaker diarization problem for every channel (we were interested in detecting the primary speaker, and categorize all other speakers as secondary). The overall diarization information can now be generated by merely combining primary speaker hypothesis from each audio channel. As seen in Fig. 1, Speech Activity Detection (SAD) is first performed to separate non-speech from speech. In this study, we used Threshold Optimized Combo-SAD (TO-Combo-SAD) for SAD [13, 12]. In the next step, the speech data is processed by the unsupervised G^3 algorithm that detects speaker change points and provides this information to the Hausdorff distance-based clustering algorithm that finds primary and secondary clusters. In what follows, we describe these algorithms in greater detail.

2.1. Unsupervised G^3 Algorithm

We propose a new method for unsupervised speaker change detection based on the work discussed in [14]. Using the theoretical foundation provided in [14], we investigated a large number of features and feature processing steps, and found a method that works well in practice for our data. We first extracted Mel-frequency cepstral coefficients(MFCCs) along with delta and delta-delta features (39-dimensions). The features were extracted for 40ms speech frames with 10ms skip rate. Additionally, a 320-dimensional real cepstrum of the linear prediction residual (RCLPR) is also used, since it models speaker-specific excitation information [15]. The 320 RCLPR features are then transformed with 51-point 1-D discrete cosine transform (DCT) to decorrelate the feature subset. Finally, the MFCCs and RCLPRs features are fused to form the final 90-dimensional fusion feature, that was used for speaker change detection.

Now, we describe the algorithm for speaker change detection. Let $\mathbf{U}$ and $\mathbf{V}$ be sets of fusion feature vectors taken from two successive 1-second time segments around time $'t'$ (we chose 1-second time window because we were interested in detecting short conversation turns, but this value can be adjusted as per application). Let $\mathbf{W} = [\mathbf{U}, \mathbf{V}]$ be the feature vectors of both frames. For detecting speaker change, we develop a binary hypothesis test, H_0 *vs.* H_1, where H_0 denotes no speaker change at time $'t'$, and H_1 denotes speaker change at time $'t'$. To facilitate the test, we build models for both hy-

potheses. On one hand, we use a 2-component GMM (Gaussian Mixture Model) to model $\mathbf{W}$. On the other hand, we use simple Gaussian function to model $\mathbf{U}$ and $\mathbf{V}$ independently. Since, one **G**MM and two **G**aussians are used in this method, we name it G^3 algorithm. The GMM parameters are estimated on-the-fly using the expectation maximization (EM) algorithm.

Now, let ϕ_w be the parameter vector of a 2-GMM estimated from $\mathbf{W}$, and let ϕ_u and ϕ_v be the Gaussian parameters for $\mathbf{U}$ and $\mathbf{V}$, respectively. If we assume the features in $\mathbf{U}$ and $\mathbf{V}$ are independent and identically distributed, we have the following expression for log likelihood L_{H_0} and L_{H_1} for both hypotheses H_0 and H_1, respectively,

$$L_{H_0} = \sum_{i=1}^{N} \log(p(u_i|\phi_w)) + \sum_{j=1}^{N} \log(p(v_i|\phi_w)), \quad (1)$$

and

$$L_{H_1} = \sum_{i=1}^{N} \log(p(u_i|\phi_u)) + \sum_{j=1}^{N} \log(p(v_i|\phi_v)), \quad (2)$$

where $p(x|\phi)$ is the likelihood of the fused feature vector x given model parameters ϕ. The detection index, D_{LLR}, is based on log-likelihood ratio (LLR) and is given by

$$D_{LLR} = L_{H_1} - L_{H_0}, \quad (3)$$

where D_{LLR} is greater than 0 whenever the 2-component GMM is a better model for the observed fused feature vector $\mathbf{W}$. Hence, speaker change (H_1) occurs when $D_{LLR} > 0$ [14].

2.2. Hausdorff distance-based Speaker Clustering

Most state-of-the-art diarization systems used for TV shows and meetings tend to use hierarchical clustering. However, research has shown that spectral clustering that involves eigendecomposition and k-means clustering is computationally simple as compared to hierarchical clustering [16]. For example, in [16], the authors used Japanese Parliament audio data that had segments of length 3 seconds or greater to compare hierarchical and spectral clustering. Spectral clustering is a global approach and hence optimal with respect to similarity criterion. On the other hand, hierarchical clustering is greedy and can lead to sub-optimal solutions. However, the performance of spectral clustering largely depends on the choice of similarity metrics.

Here, Kullback-Leibler (KL) divergence is not the best suited for audio segments of less than 3 seconds [16]. In CRSS-PLTL, short speaker turns (about 1 second) were quite common, that made it difficult to use the KL divergence metric. This motivated the need to research a more suitable metric. In this study, we propose to use Hausdorff distance as similarity measure for spectral clustering.

The Hausdorff distance assigns a scalar metric for similarity between two vectors or two matrices or a vector and matrix of different sizes. It has been found to be effective in tracking similarity among complex structures [17, 18]. Let $\mathbf{A_1}$ and $\mathbf{A_2}$ be feature matrices of dimension $m_1 \times n$ and $m_2 \times n$ where m_1 and m_2 are number of frames in both audio segments and n being the feature dimension. The Hausdorff distance between feature matrices, $\mathbf{A_1}$ and $\mathbf{A_2}$ is given as

$$d_H(\mathbf{A_1}, \mathbf{A_2}) = \max(h(\mathbf{A_1}, \mathbf{A_2}), h(\mathbf{A_2}, \mathbf{A_1})), \quad (4)$$

where $h(\mathbf{A_1}, \mathbf{A_2})$ is given by,

$$h(\mathbf{A_1}, \mathbf{A_2}) = \max_{a_1 \in \mathbf{A_1}} \min_{a_2 \in \mathbf{A_2}} \parallel a_1 - a_2 \parallel, \quad (5)$$

and $\parallel \cdot \parallel$ is some underlying norm such as L_2 or Euclidean norm on elements in $\mathbf{A_1}$ and $\mathbf{A_2}$. Here, d_H is the Hausdorff distance between two feature matrices, $\mathbf{A_1}$ and $\mathbf{A_2}$. Using Hausdorff distance as a similarity metric, various audio segments are compared and the most similar are merged together. Next, the Hausdorff distance between newly merged cluster and other clusters is recomputed and the process is repeated until we are only left with two clusters (one each for primary and secondary speakers).

2.3. Primary speaker identification

Once two clusters are identified using Hausdorff distance based clustering, primary and secondary clusters are identified in the last step. The identification can be made based on a simple observation that the primary speaker tends to be closer to the microphone than secondary speakers. This causes primary speech to be more energetic than secondary speech. We have previously exploited this fact in other studies [19, 12], and have seen that this is a fairly robust assumption that tends to get even stronger with increasing duration. By measuring the average energy in the two clusters, we assign the cluster with higher and lower energies to primary and secondary speakers, respectively. The energy computation is performed by summing the energy of the first two speech formants.

Finally, since we have multi-channel data, the energy measurements across channels can be further exploited to improve primary speaker identification. It is useful to note that while all microphones pick up every speaker's voice (due to close proximity), each speaker is loudest (most energetic) on their own microphone (owing to the physical distance separating speakers from the microphone). Additionally, it is assumed that overlapped speech is rare, and only one speaker speaks at a time (our analysis of the data showed that less than 3% of the data contained overlapped speech). In other words, there is only one primary speaker across all channels at any given time. To exploit this, we scan decisions across all channels for fixed time windows (we used 2 second windows in our experiments), and identify regions where more than one channel contains the primary speaker. For these regions, we retain primary speaker decision only for the most energetic channel, and reverse the decision to secondary speakers for other channels. This process allows us to further refine the diarization hypothesis. There were some temporal shifts in various audio streams that was not utilized in this paper.

2.4. Analysis

Once primary *vs.* secondary speaker decisions are available for each audio channel, the overall diarization information is readily made by merely combining the individual channel results. Using the basic diarization information, a number of interesting metrics can be derived for the PLTL session. In this study, we show two metrics: (i) speaker turn-taking, and (ii) speaker participation measured using speech duration.

The quality of a conversation either in a classroom scenario such as PLTL or those at workplaces can be quantified qualitatively in terms of turn-taking. More turn-taking between various speakers in a group discussions shows more engagement and hence healthy discussions. For PLTL scenario, better engagement in solving tutorial problems can conclude that students are motivated in problem solving. We used the G^3 algorithm for counting the conversational-turns taken. Total number of conversational turns taken is given by the total number of speaker-changes for each channel of PLTL. Averaging the total-turns from each channel, we get the average turns taken in PLTL session. This metric quantifies the quality of discussions in that session. We compute the speaker-changes on a sliding segment of 1 second duration. The total conversational-turns computed from various channel are summarized in Table 2.

3. Experiments

3.1. CRSS-PLTL Corpus

While collecting CRSS-PLTL corpus data, 5 PLTL teams were tracked over an entire semester. Each team consisted of 5-to-8 members, where one member was always the team leader. All teams met once every week for a total of 11 weeks, resulting in a total of 55 sessions for the corpus. All students were part of an undergraduate Chemistry course. The collection is longitudinal as it tracks individuals over a 3-month time period. Each session was 80 minutes long, and each team member wore a numbered LENA audio recording unit for the entire duration of the session. It is useful to note that the LENA digital language processor (DLP) can record audio signals for long duration upto 16 hours and has been used for a variety of human-to-human communication research studies, especially adult-child interaction [20, 21, 22].The audio data in CRSS-PLTL contains varying amounts of noise and reverberation, and at times, the noise and reverberation level can be significantly degrading. Finally, each student completed a survey after each session that sought Likert-scale ratings for subjective questions such as behavior, communication, learning, *etc.*. In order to facilitate experimental evaluation for this study, 21 minutes from one session was chosen, and manual annotations for speech activity and diarization were created. This evaluation set contained 7 parallel audio channels (corresponding to 7 team members who attended that session). We downsampled the audio data to 8 kHz before processing it. It is same for all results discussed in this paper.

3.2. Baseline System

We used the LIUM speaker diarization system as the baseline diarization system and compare its performance with the proposed system [5, 23]. The standard LIUM system was used for results presented in this paper. It is possible to use reasonable amount of labeled PLTL data for optimizing the LIUM system parameters. However, we have not optimized LIUM system for results discussed in this paper due to unavailability of enough labeled data. For all the experiments, the audio signals were downsampled at 8 kHz. The speech signal was divided into frames of size 40ms with a skip rate of 10ms. Our previous study has shown that TO-Combo-SAD worked better

Table 1: *Comparison of proposed system and LIUM baseline using Diarization Error Rate (DER) and Speech Activity Detection (SAD) Equal Error Rate (EER).*

System Used	DER (%)	EER(%)
LIUM (A)	35.80	12.54
A + TO-Combo-SAD (B)	34.20	8.67
B + Primary Speaker Identification (C)	32.76	8.67
Proposed System	**24.96**	**8.67**

than the default SAD setup in LIUM [12]. Hence, we used TO-Combo-SAD to generate speech *vs.* non-speech decisions. We constrained LIUM to 2-speaker decisions, and further used the primary speaker identification method described in Sec. 2.3 to make primary *vs.* secondary speaker decisions.

3.3. Results & Discussions

We used DER as the figure of merit for the proposed and baseline diarization systems. DER, as defined by the NIST Rich Transcription Evaluation [24], can be computed as,

$$DER = \frac{L_{fa} + L_{miss} + L_{err}}{L_{total}} \quad (6)$$

where where L_{fa} is the total number of non-speech segments detected as speech, L_{miss} is the total number of the speech segments detected as non-speech, L_{err} is the total number of speech segments that were detected as speech but clustered as incorrect speakers, and L_{total} is the total number of speech segments obtained using the ground-truth labels. Average DER across various channels was used as a metric for performance comparison. Additionally, we also compute and report equal error rate (EER) for TO-Combo-SAD. Table 1 shows DER and EER numbers for the baseline and proposed systems. Systems A, B and C are variations of the baseline LIUM system, where A is the LIUM system, B is LIUM system that takes SAD decisions from TO-Combo-SAD, C is LIUM system with TO-Combo-SAD that uses primary speaker identification described in Sec. 2.3. As seen in the table, TO-Combo-SAD (8.67% EER) delivers superior SAD decisions *vs.* LIUM SAD (12.54% EER). Furthermore, using TO-Combo-SAD and primary speaker identification reduces overall DER for the task by about 3% absolute (35.80% to 32.76%). However, the proposed diarization system is able to significantly outperform system C, and improves the DER by about 8% absolute. This is remarkable because the proposed system is unsupervised and relatively computationally inexpensive when compared to LIUM (that utilizes i-vector based solution). We believe the better performance was achieved because CRSS-PLTL data contained shorter speaker turns, where the proposed system outperformed LIUM. Further analysis of DER across each audio channel revealed that the DER for individual channels varied between 22.48% to 26.84%, that suggests stable performance. Finally, we show two analyses using the proposed system. In the first analysis, the diarization output was used to count turns taken by each student and the team leader. The speaker turns could also be estimated from the ground-truth and this was used to determine accuracy of turn taking analysis. Table 2 shows turn taking estimation performance. It can be seen that the percentage error varies between 2.7% and 7.32%, that was interesting given that DER was about 24% for this task. On average, each member took 35-to-36 turns in the 21-minute evaluation audio. Finally, we estimated how long each member spoke, by using the diarization output. In Fig. 2 (a), the proportional duration (that indicates proportional

Table 2: *Showing performance of conversational-turn taking analysis using proposed speaker diarization system.*

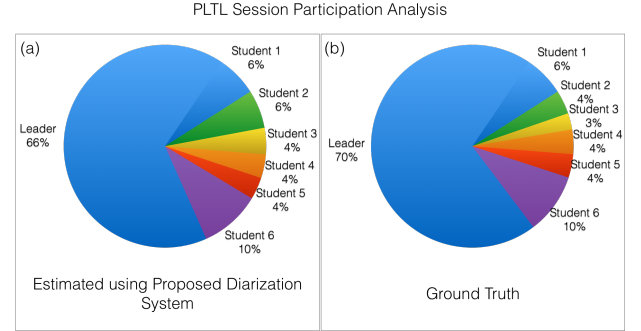| Member	Estimated Turns-taken	Error (%)
Student 1	34	5.56
Student 2	38	6.45
Student 3	27	6.86
Student 4	35	7.32
Student 5	39	4.88
Student 6	37	6.45
Leader	37	**2.70**
Mean	**35.29**	**5.75**

Figure 2: *Automatic PLTL member participation analysis using proposed diarization system and comparison to analysis generated from ground-truth labels.*

participation in conversation) is shown, and compared to a proportional participation pie chart generated using ground-truth in Fig. 2 (b). Comparison of the percentage participation numbers showed that the error were surprisingly low, and the analysis generated through proposed diarization method was rather accurate. For example, the leader occupies the conversation for almost two-thirds of the time, and students 6 and 3 contribute the most and least among students, respectively. In future work, encouraged by the results seen here, we wish to expand such analysis to the entire CRSS-PLTL corpus, and explore the ability to detect students at risk for subject material learning.

4. Conclusions

This study proposed an unsupervised speaker diarization system that used a new speaker change detection algorithm (termed unsupervised G^3 algorithm) and a new speaker clustering algorithm based on Hausdorff distance. A feature set for unsupervised G^3 algorithm that worked well for PLTL data had also been proposed. TO-Combo-SAD was used to separate speech from non-speech. The proposed diarization system was evaluated on a new corpora called CRSS-PLTL. The new corpora presents opportunity for speaker diarization research and its application in education research. In the experimental evaluations shown, the proposed diarization system significantly outperform the baseline LIUM diarization system. Finally, practical analysis using the proposed diarization system output was presented and discussed. The results and analyses presented are encouraging and motivate use of speech processing technology in studying practical problems in education research in particular, and human-to-human communication problems for small groups in general.

5. References

[1] C. C. Wamser, "Peer-led team learning in organic chemistry: Effects on student performance, success, and persistence in the course," *Journal of Chemical Education*, vol. 83, no. 10, p. 1562, 2006.

[2] K. S. Lyle and W. R. Robinson, "A statistical evaluation: Peer-led team learning in an organic chemistry course," *Journal of Chemical Education*, vol. 80, no. 2, p. 132, 2003.

[3] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization of meetings recordings beyond mfcc and tdoa features," *Speech Communication*, vol. 54, no. 1, pp. 55–67, 2012.

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

[5] A. Gallardo-Antolın, X. Anguera, and C. Wooters, "Multi-stream speaker diarization systems for the meetings domain," in *Proc. Intl Conf. Spoken Language Processing (ICSLP06), Sept*, 2006.

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

[7] S. E. Tranter, D. Reynolds *et al.*, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[8] S. H. Yella and A. Stolcke, "A comparison of neural network feature transforms for speaker diarization," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] H. Ghaemmaghami, D. Dean, and S. Sridharan, "A cluster-voting approach for speaker diarization and linking of australian broadcast news recordings," in *IEEE ICASSP*, 2015, pp. 4829–4833.

[10] A. Pikrakis, "Unsupervised audio segmentation based on restricted boltzmann machines," in *The 5th International Conference on Information, Intelligence, Systems and Applications, IISA*. IEEE, 2014, pp. 311–314.

[11] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and ngsw corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 907–919, 2006.

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

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

[14] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, 2004.

[15] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[16] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization." in *INTERSPEECH*, 2006.

[17] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[18] N. Basalto, R. Bellotti, F. De Carlo, P. Facchi, E. Pantaleo, and S. Pascazio, "Hausdorff clustering," *Phys. Rev. E*, vol. 78, p. 046112.

[19] A. Ziaei, A. Sangwan, L. Kaushik, and J. H. L. Hansen, "Prof-life-log: analysis and classification of activities in daily audio streams," in *IEEE ICASSP*, 2015, pp. 4719–4723.

[20] D. Xu, U. H. Yapanel, S. S. Gray, J. Gilkerson, J. A. Richards, and J. H. L. Hansen, "Signal processing for young child speech language development." in *WOCCI*, 2008, p. 20.

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

[22] A. Sangwan, J. H. L. Hansen, D. W. Irvin, S. Crutchfield, and C. R. Greenwood, "Studying the relationship between physical and language environments of children: Who's speaking to whom and where?" in *IEEE Signal Proc. Education Workshop 2015, Salt Lake City, Utah*, 2015, pp. 49–54.

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

[24] NIST, "Rich transcription 2004 spring meeting recognition evaluation plan," http://www.nist.gov/speech/,2004, 2004.