PROF-LIFE-LOG: PERSONAL INTERACTION ANALYSIS FOR NATURALISTIC AUDIO STREAMS

Ali Ziaei, Abhijeet Sangwan, John H.L. Hansen

Center for Robust Speech Systems(CRSS),
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A
{ali.ziaei, abhijeet.sangwan, john.hansen@utdallas.edu }

ABSTRACT

Analysis of personal audio recordings is a challenging and interesting subject. Using contemporary speech and language processing techniques, it is possible to mine personal audio recordings for a wealth of information that can be used to measure a person's engagement with their environment as well as other people. In this study, we propose an analysis system that uses personal audio recordings to automatically estimate the number of unique people and environments which encompass the total engagement within the recording. The proposed system uses speech activity detection (SAD), speaker diarization and environmental sniffing techniques, and is evaluated on naturalistic audio streams from the Prof-Life-Log corpus. We also report performance of the individual systems, and also present a combined analysis which reveals the interaction of the subject with both people and environment. Hence, this study establishes the efficacy and novelty of using contemporary speech technology for life logging applications.

Index Terms— Social Interaction Analysis, Personal Interaction Analysis, Environmental sniffing, Eigen value decomposition (EVD)

1. INTRODUCTION

Collecting personal audio recordings is becoming increasing inexpensive and feasible with the popularity of mobile personal computing devices and ubiquitous inexpensive storage. Personal audio recordings collected over an entire day contain a wealth of information pertaining to interaction with work/personal environment and people. Using speech and language processing capabilities, it is possible to mine these recordings for knowledge. For example, one could find answers to questions like "how much time did I spend driving this week?", "how much time did I spend meeting with my son helping with his homework last month?" etc. In this study, we propose a personal interaction analysis system that combines the capability of Speech Activity Detection (SAD), Speaker Diarization, and Environmental Sniffing techniques to track and record a person's activities.

To facilitate this study, we have used data from the Prof-Life-Log corpus. The Prof-Life-Log corpus contains audio recordings for an entire work day (10+ hours) collected using the LENA (Language Environment Analysis) unit [6]. The LENA unit is light and compact, and is easily worn by a person. In this study, the person wearing the LENA unit is referred to as the primary speaker. Other speakers appearing in the audio recording are referred to as secondary speakers. The personal interaction analysis system proposed is able to separate speech from background. For this purpose, we use a Speech Activity Detector. In this study, we have used an unsupervised SAD approach based on voicing measures and perceptual spectral flux proposed by Sadjadi and Hansen [1]. The unsupervised approach is suitable for our problem since the audio recordings used contain a variety of non-stationary noise-types.

Additionally, using the speech segments generated by the SAD and speaker diarization techniques, the proposed system can separate the primary from secondary speakers, and the secondary speakers from each other. This capability allows us to estimate the number of secondary speakers in the audio track, along with other conversation metrics related to primary and secondary speaker interaction (*e.g.*, turn taking behavior). A traditional diarization system is bottom-up, (*i.e.*, the process starts with segmentation and changing point detection). Subsequently, each segment is merged with other similar segments using similarity measures such as Bayesian information criteria (BIC). This process is repeated until the number of remaining nodes are equal to the number of speakers [5]. However, using BIC is time consuming and requires prior knowledge of the number of speakers. Therefore, in this study, we have proposed a new approach for speaker diarization that simultaneously segments and estimates the number of speakers using an eigen-decomposition approach.

Finally, using pause segments generated by the SAD algorithm, the proposed system uses environmental sniffing[2] techniques to classify the background audio environments into general categories such as "office", "restaurant", "car" etc. This capability allows us to track the speaker through various environments. In this study, we have used the acoustic signature vector (ASV) method recently proposed in [4] to segment and classify background audio environments. It is noted that other techniques such as [9, 10] have also been proposed for environmental detection and auditory scene analysis. The ASV technique is unsupervised which makes it more suitable for naturalistic audio streams that contain a large diverse set of audio environments.

2. PROF-LIFE-LOG

In this study, we have used data from the Prof-Life-Log corpus for evaluation. The Prof-Life-Log corpus is a collection of long single-session audio recordings in natural settings. The audio is recorded using a light-weight compact device called the LENA unit, that is capable of recording up to 16+ hours continuously. In our collection, the device is worn for the entire workday, and the audio data is captured continuously throughout the day(e.g., sample day-long recordings have included ICASSP and Interspeech Conferences). Fig. 1 shows the LENA device (attached to the shirt pocket) collecting audio data is various settings. So far, the Prof-Life-Log corpus contains 35+ days of audio recordings, resulting in a total collection of 300+ hours. We have annotated speaker and environment information in the corpus. For environment sniffing, we have annotated approximately 5 hours of data for use in evaluation. For the speaker diarization task, we have annotated 6 hours of Prof-Life-Log containing 10

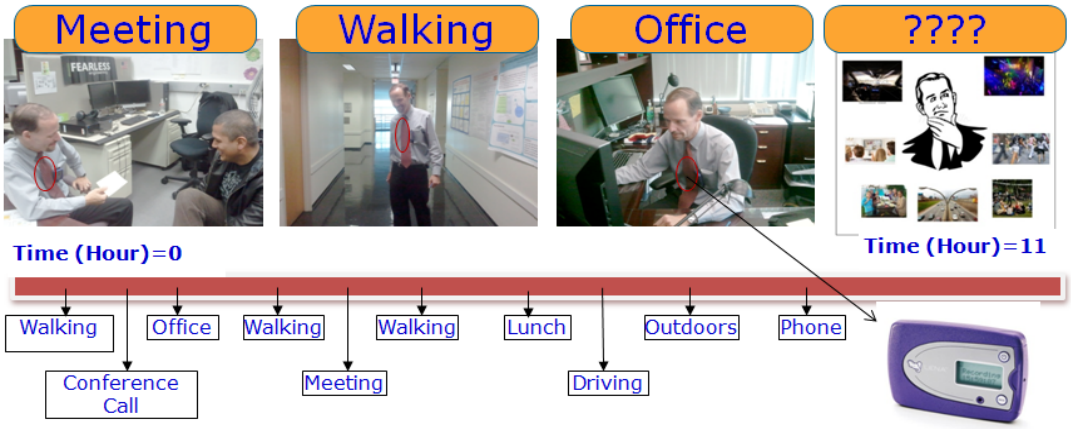

Fig. 1. Data collection using the LENA unit: A single session consists of 10+ hours of audio recording with the speaker constantly carrying the unit. Speech is collected in a wide variety of backgrounds such as Cafeteria, Office, Meeting, Walking, Driving, etc. The primary speaker wears the LENA unit, and his/her interaction with secondary speakers is captured.

speakers.

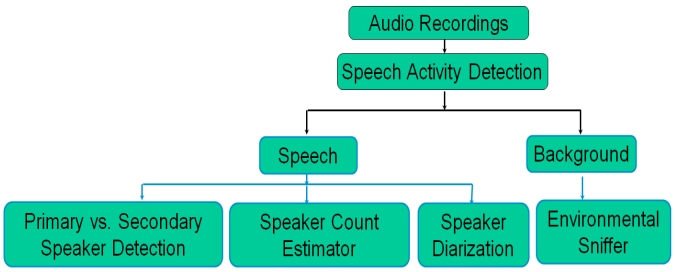

Fig. 2. Proposed personal interaction system uses speech activity detection (SAD), speaker diarization and environmental sniffing techniques to analyze user interaction with people and environment.

3. PROPOSED SYSTEM

As shown in Fig. 2, our system consists of five different components, namely, Speech Activity Detection (SAD), Environmental Sniffer, Primary vs. Secondary Speaker Detector, Speaker Count Estimator and Speaker Diarization system. In what follows, we describe each component.

3.1. Speech Activity Detection (SAD)

The incoming audio recordings are first segmented into speech and background using the SAD proposed in [1]. The speech signal is first segmented into 10ms frames and then energy and spectral flux features are computed for each frame. Subsequently, a 2-mixture GMM is trained for both features using the entire audio file. In order to assign each frame to speech or background, a threshold is first set to the average of the mixture means. Subsequently, speech/background decisions are made for each frame by comparing the feature value to the threshold.

The raw SAD output requires smoothing as it ocassionally contains rapid fluctuations between speech and pause decisions. For smoothing, we use majority voting within a 1 second window, (*i.e.*, the 1s segment is declared speech as it contains more speech decisions and vice-versa). The post smoothing decisions are the final SAD output and are used in the remaining sub-tasks.

3.2. Speech Analysis

3.2.1. Speaker Count Estimation

We propose a new approach for speaker diarization that simultaneously estimates speaker count and segments. Traditional diarization systems use BIC (Bayesian Information Criterion) to merge similar clusters. This process is computationally expensive and inefficient for large audio recordings. In this study, we use k-means clustering followed by eigen-analysis to perform segmentation and speaker count estimation. This process is described below.

Using the speech segments from SAD, we first train a GMM (conceptually similar to a Universal Background Model used in speaker recognition). Let X_s be the acoustic feature vectors for segment s that is used to train the mentioned GMM. Next, we adapt the GMM using each speech segment. Let the model obtained by adapting the GMM using the j^{th} segment be M_j. Finally, let

$$P(X_k|M_j) \tag{1}$$

be the posterior probability of segment X_k being generated by Model M_j. Let the k^{th} speech segment be denoted by

$$V_k = [P(X_k|M_1), P(X_k|M_2), ..., P(X_k|M_M)]^T. \tag{2}$$

Using these vectors, we construct the correlation matrix

$$M_{Corr} = [V_{1n}, V_{2n}, ..., V_{Mn}]^T * [V_{1n}, V_{2n}, ..., V_{Mn}], \tag{3}$$

where V_{in} is the normalized version of V_i (*i.e.*, the norms equals 1). Now, by performing eigen-decomposition of M_{Corr},

$$M_{Corr} = U * \sum * U', \tag{4}$$

we can count the number of eigenvalues in $\sum$ larger than a predefined threshold set using the p-value test [11]. Using this threshold, we can determine the number of eigenvalues greater than the p-value test threshold, and we assume that this number is equal to number of speakers in the audio file.

To implement the p-value test, we process all segments through our system 500 times and for each, perform eigen decomposition over the correlation matrix described earlier. By fitting a Gaussian distribution to the derivative of the curves fitted to eigen values in each iteration, we compute the bottom 5 percent value in that distribution and set this as the threshold.

3.2.2. Speaker Diarization

Using the speaker count estimate, we perform k-means clustering on vectors $V_{in}, i = 1, ..., M$ to obtain speaker clusters [7]. The k-means algorithm provides a locally optimal solution and is highly sensitive to initial values. Therefore, we propose an eigenvector based strategy to select the initial set of vectors as opposed to a randomly selected vector set. In this approach, we decompose M_{Corr} as follows,

$$\begin{aligned} M_{Corr} &= U * \sum^{1/2} * \sum^{1/2} * U^T \\ &= (\sum^{1/2} * U^T)^T * (\sum^{1/2} * U^T) \\ &= V_{New}^T * V_{New}. \end{aligned} \quad (5)$$

Next, we use the first K vectors of V_{New} as the initial vector set for k-means clustering.

3.2.3. Primary vs. Secondary speaker detection

The primary speaker's speech in our audio recordings is easily separated from secondary speakers since the primary speaker is closest to the microphone. It is reasonable to assume that this characteristic would be generally true across personal audio recordings (due to the nature of data collection). Therefore, in the proposed system, we first separate primary from secondary speakers using the proposed diarization scheme (and fix the speaker count to two). Subsequently, we apply the speaker diarization approach again on secondary speakers to estimate the unknown speaker count and to obtain individual speaker segments.

3.3. Environmental Sniffing

Using the background segments generated by SAD, we use environmental sniffing[4] to track the audio background environments of the primary speaker. In this study, we have used the Acoustic Signature Vector (ASV) technique proposed in [4] for environment segmentation.

The ASV extraction process in briefly reviewed here. First, a GMM (Gaussian Mixture Model) is trained using large quantities of diverse audio material and used as the background acoustic model. Next, the ASV is computed as follows. Let X_e be the acoustic feature vector (such as MFCCs) that is used to train M-mixture GMM, where m_j is the j^{th} mixture. Finally, let $P(X_e|m_j)$ be the posterior probability of feature vector X_e being generated by mixture m_j. Now, let the k^{th} audio segment be denoted by V_k. Assuming V_k contains N feature vectors,(*i.e.*, $V_k = [X_{e1}X_{e2}\ldots X_{eN}]$), we compute the average posterior probability of mixture m_j across all feature vectors in V_k as,

$$q_m = \frac{1}{N}\Sigma_{i=1}^{N} P(X_{ei}|m_j). \quad (6)$$

Next, we construct the a posterior probability vector Q as,

$$Q = [q_1 q_2 \ldots q_M]^T, \quad (7)$$

and term Q as the acoustic signature vector (ASV). Let the ASV for the V_k (k^{th} segment) be denoted by Q_k.

In the next step, the ASVs are clustered into homogenous groups using the k-means algorithm. In order to determine the number of unique environments in the audio file, we use the proposed eigen-decomposition based speaker count estimation technique (see Sec. 3.2.1). The only modification being that we estimate the number of environments (as opposed to number of speakers).

4. RESULTS

In this study, we have used MFCC (Mel-frequency cepstral coefficient) features for the speaker diarization and environmental sniffing tasks. The MFCCs were extracted with a frame duration of 25ms and 10ms skip using 27-filterbanks. We used 12 static coefficients with energy. For speaker diarization and environmental sniffing, we used 64 mixture and 128 mixture GMM, respectively.

Fig. 4 shows the DET (detection error trade off) curve for the SAD employed. The miss rate ($Miss - SAD$) is the percentage of speech frames that were mis-classified as background, and false-alarm rate ($FA - SAD$) is the percentage of background frames mis-classified as speech frames. For this study, we choose an operating point that provides 9% miss-rate and 2.1% false-alarm rate. The reported SAD results are very competitive given the naturalistic variable of audio backgrounds in Prof-Life-Log.

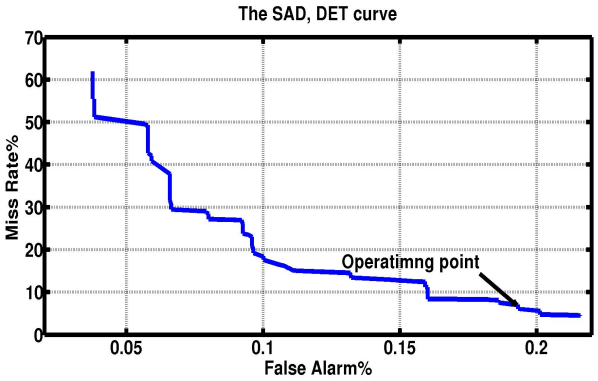

Fig. 4. DET curve showing SAD performance:Prof-Life-Log data

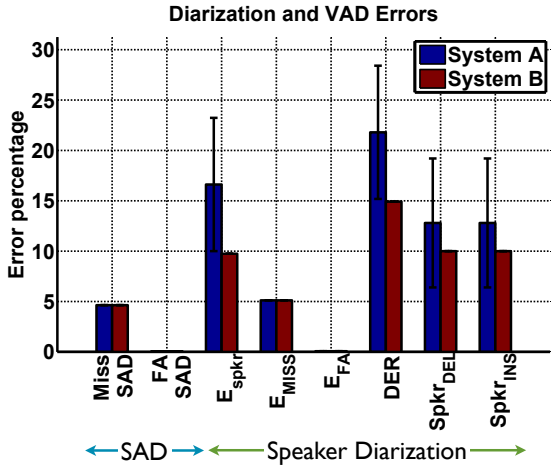

Fig. 5. Speaker Diarization Performance using standard NIST metrics:System A (Random Initialization), System B (Eigen-vector based Initialization)

For speaker diarization, we evaluated two variations of the proposed system, namely, System A: uses random initialization for k-means clustering, and System B: uses proposed eigenvector based initialization for k-means clustering (see Sec. 3.2.2). Figure 5 shows the speaker diarization results for Systems A and B. The first two evaluation metrics $Miss - SAD$ and $FA - SAD$ represent the SAD performance. In this study, we have used the standard NIST speaker diarization evaluation package [8]. Here, we report

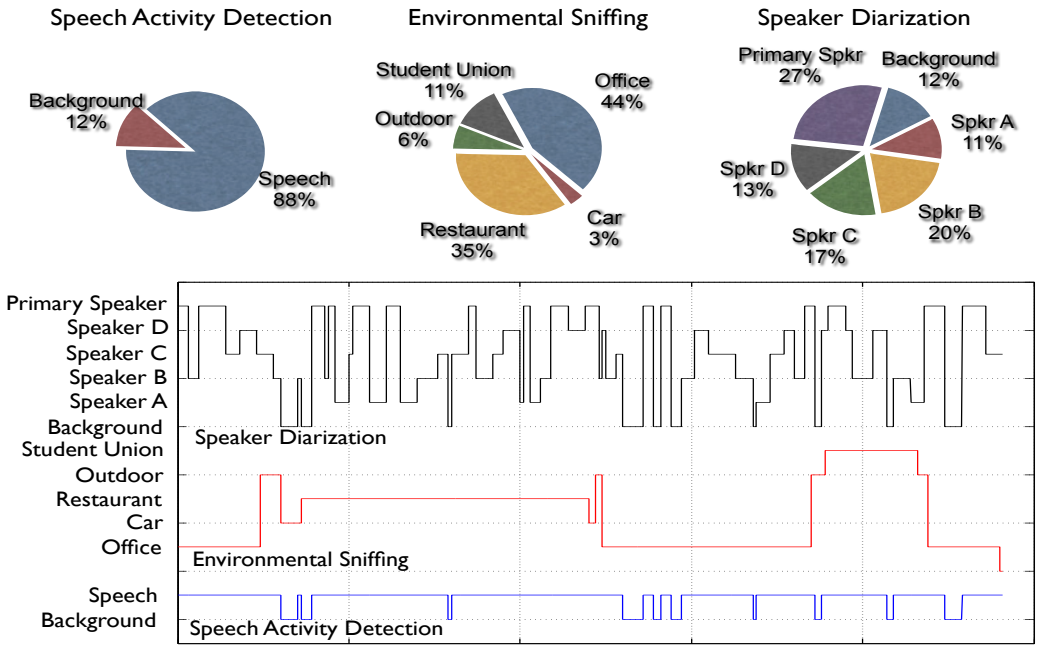

Fig. 3. Personal Audio Recordings Analysis using SAD, Speaker Diarization and Environmental Sniffing.

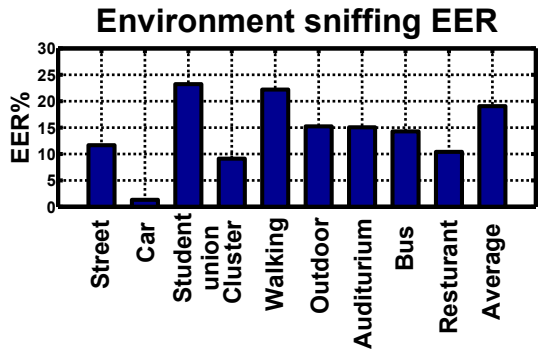

Fig. 6. Environment sniffer performance: Equal Error Rate (EER) for various Environments.

the standard NIST metrics in Fig. 5: (i) speaker error (E_{spur}), (ii) missed speech (E_{MISS}), (iii) false alarm speech (E_{FA}), and (iv) diarization error rate (DER). The final two metrics, speaker insertion ($Spkr_{INS}$) and deletion ($Spkr_{DEL}$) errors represent the numbers of speakers falsely hypothesized and missed by the diarization system.

For system A, we repeated the diarization task 100 times with different initializations for the k-means algorithm. Fig. 5 shows the mean and the first-standard-deviation of the diarization results for System A. The sensitivity of system A to random initialization can be observed from the figure, as the DER ranges from 15% to 28% depending upon initialization. Furthermore, the merit of eigenvector based k-means initialization can also be observed at system B performs equal to or better than system A.

Fig.6 shows the performance of the Environmental Sniffer in terms of EER (Equal Error Rate) for different environments. The average performance across all environments is 19% EER. The best and worst EERs are obtained for car (2%) and student union(Cafeteria)(23%) environments.

Fig.3 shows the combination of SAD, speaker diarization and environmental sniffing results over approximately 3+ hours of continuous Prof-Life-Log data. The analysis reveals that primary speaker was initially in his office. While in office the speaker has brief conversations with speakers C and D. Subsequently, he drove to a restaurant for lunch with speakers C and D. The analysis also suggests that the primary speaker had lunch with all secondary speakers. Hence, it can be seen that the timeline presented in Fig. 3 summarizes the main events of the day and reveals speaker location and movement during the day.

5. CONCLUSION

In this study, we have shown the efficacy of using SAD, speaker diarization and environmental sniffing in building a personal audio analysis system. We have evaluated the proposed system using Prof-Life-Log corpus which contains naturalistic audio streams. In particular, we have demonstrated that it is possible to perform speech/pause detection, speaker count estimation, speaker diarization, primary and secondary speaker detection and environment classification with good accuracy. We have also demonstrated that it is possible to perform interesting analysis relating to subject movement and engagement by combining the outputs of SAD, speaker diarization and environmental sniffing systems.

6. RELATION TO PRIOR WORK

Speech activity detection (SAD), speaker diarization and environmental sniffing are well established areas of research [1, 5, 4, 2]. In this study, we propose a personal interaction analysis system that combines the capability of the mentioned techniques to deliver novel analysis capability. We have also proposed improvements to speaker diarization and environmental sniffing techniques in order to build more efficient systems for naturalistic audio streams such as Prof-Life-Log with extended duration up to 16 ours.

7. REFERENCES

[1] Sadjadi S.O., Hansen J.H.L., unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux, submitted to IEEE Signal Processing Letters, Sept. 2012.

[2] Akbacak M., Hansen, J.H.L. ," Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems ",IEEE Trans. on Audio, Speech, and Language Proc.,Vol. 15, Pages:465-477, Feb. 2007.

[3] Sangwan A., Ziaei A. and Hansen J.H.L,"ProfLifeLog: Environmental Analysis and Keyword Recognition for Naturalistic Daily Audio Streams",ICASSP 2012, Kyoto, Japan, May 25-30.

[4] Ziaei A., Sangwan A., Hansen J. H.L., " Prof-Life-Log: Audio Environment Detection for Naturalistic Audio Streams ", Interspeech 2013, Portland, OR, 2012.

[5] Anguera M., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., " Speaker Diarization: A Review of Recent Research ",IEEE Trans. on Audio, Speech, and Language Proc.,Vol. 20, Pages:356-370, Feb. 2012

[6] http://www.lenafoundation.org/ProSystem/Overview.aspx

[7] Shum S., Dehak N., Chuangsuwanich E., Reynolds D., and Glass J.,"Exploiting Intra-Conversation Variability for Speaker Diarization", Interspeech 2011 Florence, Italy, Aug.

[8] NIST, Diarization error rate (der) scoring code, 2006, www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl.

[9] Ellis, D.P.W. and Lee, K., "Minimal-impact audio-based personal archives," Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences, ACM, pp. 39–47, 2004.

[10] Shao, Y. and Srinivasan, S. and Jin, Z. and Wang, D.L., "A computational auditory scene analysis system for speech segregation and robust speech recognition," Computer Speech and Language, Elsevier, vol. 24, no. 1, pp. 77–93, 2010.

[11] Schervish, M. J., "P Values: What They Are and What They Are Not," The American Statistician, vol. 50, no. 3, pp. 203–206, 1996.