PROF-LIFE-LOG: ANALYSIS AND CLASSIFICATION OF ACTIVITIES IN DAILY AUDIO STREAMS

Ali Ziaei, Abhijeet Sangwan, Lakshmish Kaushik and 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, lakshmish.kaushik, john.hansen}@utdallas.edu

ABSTRACT

A new method to analyze and classify daily activities in personal audio recordings (PARs) is presented. The method employs speech activity detection (SAD) and speaker diarization systems to provide high level semantic segmentation of the audio file. Subsequently, a number of audio, speech and lexical features are computed in order to characterize events in daily audio streams. The features are selected to capture the statistical properties of conversations, topics and turn-taking behavior, which creates a classification space that allows us to capture the differences in interactions. The proposed system is evaluated on 9 days of data from Prof-Life-Log corpus, which contains naturalistic long duration audio recordings (each file is collected continuously and lasts between 8-to-16 hours). Our experimental results show that the proposed system achieves good classification accuracy on a difficult real-world dataset.

Index Terms— Long Duration Personal Audio Recordings, Audio Analysis, Daily Summarization, LENA

1. INTRODUCTION

Long duration personal audio recordings (PARs) capture an individual's daily activities and interactions in rich details. In Prof-Life-Log corpus, the subject wears an audio recording device (called LENA [1]) which captures uninterrupted continuous audio recordings that can last between 8-to-16 hours. In this manner, the collection captures a significant proportion of the subject's daily human experience. It is interesting to consider that large collections of daily PARs that contain several weeks, months or years of data should contain sufficient information to start commenting on various aspects of the individual, such as personality, likes and dislikes, aspirations, productivity, *etc.*. Researching speech and language processing techniques to develop techniques that can offer insights into our own lives is challenging and intriguing.

In the past, researchers have attempted to develop techniques to automatically detect and retrieve a variety of information from long duration audio recordings. The ability to automatically analyze background audio or environment has received most attention. In [2], the authors propose to develop retrieval tools for PARs, and demonstrated the ability of cluster audio in broad acoustic events. In another work [3], a fast audio fingerprinting method was proposed that can automatically search for repeat acoustic events, such as ringtones, *etc.*. While the audio background in PARs is interesting and rich in information, it has been argued that speech contains more useful information [4]. In [4], the authors note the importance of good quality speech activity detection (SAD) for long duration PARs in order to extract the speech signal reliably in diverse, dynamic and often noisy acoustic backgrounds. In [5], the authors demonstrated the inability of conventional unsupervised SADs to deal with peculiar acoustic conditions of long duration audio such as extended non-speech or sparse-speech periods, and proposed a new solution to mitigate this common problem in PARs.

More recently, we have pursued work to investigate the feasibility of employing speech and speaker recognition techniques in long duration PARs. In [6], a system that automatically characterizes background environments was proposed and then exploited to drive better keyword recognition performance. This work was extended in [7], where speech recognition, speaker diarization and environment recognition systems were combined to build a system that could reveal details of the subject's daily interaction with both people and environment. Finally, in [8], a new technique that can count the total number of words spoken in a day for long duration PARs was proposed. The proposed solution was shown to be extremely effective on typical workdays.

In this study, a new method to analyze daily activities in PARs is presented. In Prof-Life-Log, a typical workday consists of various events, which in general fall into one of two categories: (i) some type of interaction involving multiple individuals (e.g., meetings, classroom, seminars, conference calls, etc.) or (ii) some form of alone time (reading, typing, thinking, driving, etc.) where the subject is by himself. The proposed method allows us to classify the mentioned events. This capability allows us to provide a high level semantic segmentation of the day. Furthermore, the approach can also inspect the details of each event to analyze and understand,

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

identify anomalies or outliers, *etc.*

The new method exploits a number of audio, speech and lexical features to characterize events in daily audio streams. Specifically, we employ features that capture the statistical properties of conversations and turn-taking behavior, which allows us to capture the differences in interactions. For example, seminars and classrooms tend to be more monologue-like (one person dominates although multiple participants are present), and meetings tend to be more dialogue-like (participants share time and one person is less likely to dominate). Additionally, we employ lexical features that allow us to track topics. For example, topic detection would allow us to discriminate between administrative and research meetings, even though the style of interaction may be similar for both these types of events. In this study, we evaluate the proposed system on 9 days of audio data. Our experimental results show good ability to identify events in daily audio streams.

2. SYSTEM DESCRIPTION

The proposed system consists of three major parts, audio pre-processing, feature extraction and classification. The workflow diagram of the system is shown in Fig. 1.

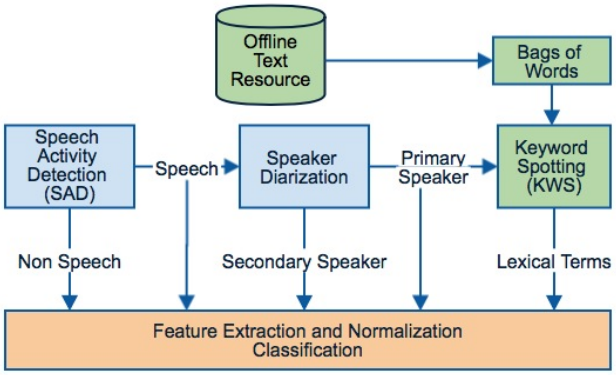

Fig. 1. Block diagram of proposed activity detection system.

2.1. Audio Pre-processing

Prof-Life-Log audio contains a variety of conversations in diverse acoustic environments. Therefore, Speech Activity Detection (SAD) is an important system component as it allows us to consistently segregate speech from various time-varying, noisy acoustic backgrounds. Once speech has been successfully separated from background, we need to separate primary from secondary speech (in Prof-Life-Log, the subject who wears the device is referred to as the primary speaker and all others are collectively referred to as secondary speaker). We use speaker diarization technology to separate primary from secondary speech.

2.1.1. Speech Activity Detection

In this study, we use the TO-Combo-SAD (Threshold Optimized Combo SAD) algorithm for separating speech from noise. In long durations PARs, extended periods of non-speech or sparse-speech is very common. In a previous study [5], we have shown that contemporary state-of-the-art SAD techniques such as Combo-SAD which have shown good performance in severe environments, fail to address both these problems and results in large number of false-alarms. Additionally, we had also shown that the TO-Combo-SAD is able to address both the mentioned problems and deliver significantly improved results. In our previous experimental evaluation, we saw best performance for TO-Combo-SAD when the threshold parameter α was set to 0.4 [5], and this was the operating parameters for this study as well. Interested readers can refer to [5] for algorithm details.

2.1.2. Speaker Diarization

Primary *vs.* secondary speaker detection is a special form of speaker diarization. In general, the primary speaker (subject wearing the audio recorder) is louder than secondary speaker(s) because of the closer proximity to the device microphone. Intuitively, primary speech energy should be greater than secondary speech, and we exploit this phenomenon in our solution. It is noted that we expect this phenomenon to generalize across recording devices.

In this study, we first employ the open source speaker diarization toolbox by LIUM [9] to provide initial diarization. The output of the LIUM toolbox provides an initial hypothesis of the first and second speaker. In the next step, we compute energy for the hypothesized segments. By averaging the segment-level energy estimates for first and second speakers, and then selecting the speaker with higher energy level as primary speaker, we can achieve primary *vs.* secondary speaker separation.

2.2. Features

The pre-processing step provides us with basic classification where the audio signal is now segmented into acoustic background, primary and secondary speech. This basic classification is still able to provide information about when the subject is by himself or engaged in a conversation. It also reveals information about turning taking behavior in conversations. By better characterizing these aspects of the audio signal, it is possible to start segregating various events in daily recordings. Previous studies have shown that high level features that attempt to capture attributes of conversations can be useful predictors of behavior. For example, in [10], high level conversational features showed strong relation with cohesion in meetings.

Towards this, we propose to extract 14 different features that capture various attributes of the underlying event in the audio signal. These features attempt to extract information pertaining to acoustic background, speech and vocabulary of the audio signal. The entire audio stream is divided into con-

tiguous 5-minute audio blocks and the features described below are extracted for every 5-minute block.

2.2.1. Audio Features

- **Total duration of non-speech segments** divided by **Audio duration**: High values of this feature indicate that the audio block is dominated by pause (or speech is absent or sparse which is likely in alone time).

2.2.2. Speech Features

- **Total duration of speech segments** divided by **Audio duration**: High values indicate speech is dense (which is very likely in conversations).
- **Mean duration of speech segments** divided by **Audio duration**: High values indicate single speaker dominating (more likely in seminars, classroom lecture, *etc.*).
- **Total duration of primary speech segments** divided by **Audio duration**: High values indicate that the primary speaker dominates the conversation.
- **Mean duration of primary speech segments** divided by **Audio duration**
- **Total duration of secondary speech segments** divided by **Total duration of speech segment**: High values indicate that the primary speaker does not dominate the conversation.
- **Mean duration of secondary speech segments** divided by **Total duration of speech segment**
- **Total duration of secondary speech segments** divided by **Audio duration**
- **Mean duration of secondary speech segments** divided by **Audio duration**
- **Mean duration of turn to pause** divided by **Audio duration**
- NIST Signal to Noise Ratio**(NIST STNR algorithm)**
- WADA Signal to Noise Ratio**(WADA)** [11].

2.2.3. Lexical Feature

It is possible to make general distinctions between broad event classes using speech-pause and conversational turn-taking characteristics. However, further classification requires tracking user vocabulary or conversation topic. For example, an administrative meeting is easily differentiated from research meeting via lexical terms, but is nearly impossible to do based on conversation turn-taking behaviour alone. In this study, we are particularly interested in identifying research discussions and separating such events from all others.

In order to accomplish this purpose, we first build a bag-of-words that is focussed on identifying keywords that are strongly tied to research discussions. Next, we employ keyword spotting to detect the presence of such keywords in the audio file. Finally, we compute the density of research terms in the discussion by tracking the ratio of research terms used divided by total number of words spoken. We used the system described in [8] to compute total number of words spoken.

Intuitively, research conversations would be dense in research keywords (coming from the bag-of-words described above), and vice-versa. In this study, we construct the bag-of-words by parsing all research publications from 3 Interspeech conferences (a total of 2211 papers and 4,277,494 words). Using a part-of-speech tagger, we identified all noun terms from the papers along with their frequency-of-occurrence, and these were chosen as candidate terms for the bag-of-words [12]. Next, we measured the frequency-of-occurrence of these terms in general English background text (captured from various sources and in general domain independent data). Finally, using the frequency-of-occurrence measurements, we choose terms that are very frequent in research text and infrequent in background text (which is basically the term frequency-inverse document frequency method). We chose the top N terms that met the mentioned criterion.

For speech recognition, we employed a Kaldi-based medium vocabulary recognizer [13]. The language model used for recognition included the bag-of-words terms (and hence all keywords were in vocabulary). The acoustic model was trained using a mix-style approach where conversational speech from multiple sources were used (data from Prof-Life-Log was not used for training). The audio data used for training was degraded using various noise-types and SNRs (signal to noise ratios) in order to provide required noise robustness for Prof-Life-Log data. The output of the speech recognition process was a word lattice. A finite state transducer (FST) based method was used to search the word lattices for keywords [14]. In parallel, the word lattices were converted into phone lattices, and the PCN-KWS (phone confusion network keyword spotting) method was employed to search for keywords as well [15]. Subsequently, the search results from the two methods were combined (by simple likelihood combination) to yield the final keyword result list.

2.2.4. Feature Processing and Classification

The mentioned features are first mean and variance normalized. Next, PCA (principal component analysis) is performed on the features for dimension reduction. Finally, the reduced dimension features are supplied to a multi-class support vector machine (SVM) with radial basis function (RBF) kernel for model training and evaluation.

3. DATA

In this study, we use 9 days data from Prof-Life-Log corpus [6]. In order to support evaluation, the 9 days data was annotated for events. The audio files was first segmented into 5-minute contiguous audio blocks, and subsequently annotated for events at the audio block level. Annotators were asked to label each audio block as faculty-meeting, research-meeting, staff-meeting, self-study and conference call. The subject's

outlook calendar for the annotated days was made available to the annotaters for guidance.

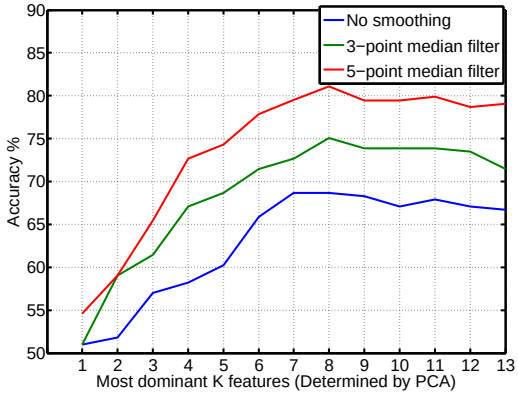

Fig. 2. Performance accuracy for the proposed system.

4. RESULTS AND DISCUSSION

For system evaluation, we performed 5-fold validation by splitting the dataset described in previous section into 80% for training and 20% for evaluation. The performance results reported here are averaged over the 5 trials.

In the first experiment, we analyze the impact of dimension reduction on classification due to PCA. By varying the principal components used for classification from 1-to-13, we compute the corresponding values for performance accuracy on the 5-way classification task. Since the underlying events are unlikely to vary fast with time, we imposed temporal constraints on the SVM output decisions by employing median filtering (the SVM decisions are first chronologically ordered). Altogether, we measured the impact of applying 3, 5, and 7 window median filter on raw SVM outputs.

The performance results for this experiment are captured in Fig.2. From the figure, it is observed that the classification performance first steadily increases as the PCA dimensions used for classification are increased, then plateaus out, and finally decreases slightly. The best performance is seen for the first 8 dimensions. Additionally, this trend is seen for all variants of temporal constraints that we applied on the data. Finally, it can be seen that the 5-point median filter seemed to work best, and corresponds to an overall accuracy of about 82%. The classification confusion matrix for the system with the best configuration is shown in Table 2. From the table, it is seen that the best and worst performance is obtained for conference calls and staff meetings, respectively.

In the next experiment, we are interested in comparing different evaluation days with each other. First, we used k-means clustering to find 32 cluster centers in the 8-dimensional feature space (as mentioned previously). Next, we assigned each 8-dimensional vector to the closest cluster center (using simple distance measure). In this manner, we obtained the cluster membership for all observations. Using this method, we generate cluster membership counts for every day, and normalize the count by the total number of observations for that day. This process yields a single 32-dimensional vector for every day, which provides a compact snapshot of the daily activity. Using this vector representation, we compute the cosine distance between all the 36 day pairs and these are shown in table 4.

Table 1. Confusion matrix for 5 activities with 5-point median filter (i.e., **(FM): Faculty meeting, (RM): Research meeting, (SM): Staff meeting, (SS): Alone Time (AT) and (CC): Conference call**, respectively.

	FM	RM	SM	AT	CC
FM	**79.91**	9.65	10.41	0.01	0
RM	19.62	**73.22**	6.51	0.03	0.62
SM	21.67	9.39	**62.78**	6.15	0
AT	10.00	6.22	10.89	**72.90**	0
CC	2.68	12.08	0.99	0	**84.25**

From the table, we can see that days 2-and-5 are most similar, and days 7-and-8 are least similar. Additionally, day-1 was observed to be least similar to any other day, and day-4 the most similar to other days (based on average cosine distance with all other days). In general, days 3, 4 and 7 seemed to form one group (as days dominated by faculty and staff meetings). Also, days 2, 5, and 8 group together (as days dominated by staff meeting and alone time). For days 6 and 9, all activities seemed to be present in similar proportions. Day 1 was an outlier as it alone contained a conference call.

Table 2. Comparing Days using Daily Activity Profile

	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1	1	.27	.24	.40	.54	.58	.26	.34	.58
D2	.27	1	.25	.56	**.81**	.39	.28	.78	.50
D3	.24	.25	1	0.73	.21	.51	**.80**	**.18**	.54
D4	.40	.56	.73	1	.55	.77	.77	.50	.82
D5	.54	.81	.21	.55	1	.52	.22	**.80**	.57
D6	.58	.39	.51	.77	.52	1	.68	.39	.78
D7	.26	.28	.80	.77	.22	.68	1	**.16**	.64
D8	.34	.78	.18	.50	.80	.39	.16	1	.44
D9	.58	.50	.54	.82	.57	.78	.64	.44	1
Avg	.47	.54	.5	.68	.59	.62	.53	.51	.65

5. CONCLUSION

A new method to analyze and classify daily activities in personal audio recordings (PARs) has been presented. The new method uses speech activity detection (SAD), speaker diarization, and a number of audio, speech and lexical features to characterize events in daily audio streams. The proposed system was evaluated on 9 days of data from Prof-Life-Log corpus, and an overall classification accuracy of approximately 82% was obtained. Additionally, a new method of analyzing daily activities using daily a different days was shown. In the future, we are interested in expanding the scope of the experiment to include several tens of days, and further research analysis techniques that allow us to organize, compare, and cluster days in the Prof-Life-Log corpus.

6. REFERENCES

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

[2] D. Ellis and K. Lee, "Accessing minimal-impact personal audio archives," *IEEE Multimedia*, vol. 13, no. 4, pp. 30–38, 2006.

[3] J. Ogle and D. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*. IEEE, 2007, vol. 1, pp. I–233.

[4] K. Lee and D. Ellis, "Voice activity detection in personal audio recordings using autocorrelogram compensation," in *INTERSPEECH 2006: ICSLP: Proceedings of the Ninth International Conference on Spoken Language Processing: September 17-21, 2006, Pittsburgh, Pennsylvania, USA*. ISCA, 2006, pp. 1970–1973.

[5] 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*, 2014.

[6] A. Sangwan, A. Ziaei, and J.H.L. Hansen, "Proflifelog: Environmental analysis and keyword recognition for naturalistic daily audio streams," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4941–4944.

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

[8] A. Ziaei, A. Sangwan, and J.H.L Hansen, "A speech system for estimating daily word counts," in *Interspeech 2014*, 2014.

[9] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," Tech. Rep., Idiap, 2013.

[10] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 563–575, 2010.

[11] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis.," in *INTERSPEECH*, 2008, pp. 2598–2601.

[12] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*, " O'Reilly Media, Inc.", 2009.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[14] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[15] A. Sangwan and J.H.L Hansen, "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, 2010, pp. 711–715.