# Prof-Life-Log: Audio Environment Detection 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

In this study, we develop a new system for real world audio environment matching. Environment detection within unknown audio streams requires a system that operates in an unsupervised manner since it will be faced with unknown environments without prior information. In addition, the overall solution should be computationally efficient for large audio collection. In the proposed approach, a Gaussian mixture model(GMM) is trained on large amounts of unlabeled audio data and used as a background acoustic model. Subsequently, an acoustic signature vector (ASV) is computed for each environment. Here, the ASV vector is designed to capture the unique acoustic characteristics of an environment. Using the ASV vectors, we demonstrate that it is possible to compute an effective similarity measure between two acoustic environments. We demonstrate the performance of the proposed system on real-world audio data, and compare it to a traditional GMM-UBM (Universal Background Model) system. Experiments show that our system achieves an equal error rate (EER) that is $+35\%$ better than a baseline GMM-UBM system.

**Index Terms**: Audio Environment Detection, Acoustic Signature, Real word audio data, Prof-Life-Log

## 1. Introduction

Audio environment detection, classification and categorization in personal audio recordings where an individual's entire day is collected as a single session is very challenging and interesting. Collecting personal audio recordings is becoming increasing inexpensive and feasible with the advent of mobile personal computing devices (such as smartphones) and ubiquitous inexpensive storage (such as cloud) [1]. The capability of audio environment detection would be the backbone of applications that would allow users to search through their audio (and video) histories, generate audio-environment summaries, etc.

Audio environment detection solutions comprise of two main parts, the system (algorithm and modeling) used for audio environment matching and the set of acoustic features. A number of features have been researched for this purpose, namely, zero crossing rate(ZCR), Mel-frequency cepstral coefficients(MFCCs), band-energy, spectral centroid, bandwidth, spectral roll-off, and spectral flux [2, 3]. Additionally, the system can be any combination of these categories including; supervised, unsupervised, feature based, or semantic based techniques. In [4], the authors compare different feature based supervised system based on k-nearest neighbor, hidden Markov model(HMM) and discriminative HMM. Also, [5] proposed an unsupervised feature based system for audio classification and segmentation in which the system divides the audio file into the homogenous parts based on $T^2$-$BIC$ [6] along with a clustering process into similar parts to partition classes using the GMM. On the other hand, instead of using just features, we can extract audio environments and focus measuring their similarity [7]. Extracting audio elements can be accomplished in a supervised or unsupervised manner. In a supervised approach, one model per environment is trained using training data a prior and used for audio environment detection during test[8]. In [9], an unsupervised system was proposed where each audio file was mapped into a latent space using a singular value decomposition(SVD), and subsequently compared to reference templates using these vectors.

Traditionally, audio environment detection has been well explored for broadcast news and meetings datasets for applications such as information retrieval, surveillance, knowledge discovery, etc. [10, 8, 11]. Here, the research work has focussed on identifying a small set of acoustic environments in relatively well controlled environments with high-quality recordings. In contrast, personal audio recordings can contain very diverse acoustic environments that can change rapidly (e.g., walking from a quiet office to a noisy cafeteria). In such scenarios, techniques that employ supervised model building with predetermined labels are rendered ineffective as they cannot address the open-set dynamic nature of the problem. Therefore, in this study, we propose an unsupervised technique that follows a query-by-example paradigm. By providing an example recording as a query, the entire collection can be searched for similar environments. This allows the proposed solution to be versatile and handle new previously unseen environments. The proposed algorithm is also designed to be fast and efficient, in order to process large quantity of data quickly.

To facilitate this study, we have been collecting audio-recordings using a portable audio recording unit called LENA (Language Environment Analysis) [12]. Till date, we have collected 35+ recordings where each recording session lasts about 10+ hours (which is a typical workday). The LENA unit is light and compact, and is easily worn by a person. In our collection, the unit is worn for the entire day and it is continuously collecting data. This collection (known as the Prof-Life-Log corpus) provides an unique and unprecedented opportunity to explore audio environment detection for real-world naturalistic audio streams.

In this study, the proposed audio detection approach is evaluated on two datasets derived from the Prof-Life-Log corpus. The first dataset represents a controlled collection with homogenous recordings of various environments. The second dataset represents a real-world naturalistic collection. We compare and contrast the environment detection results for these two datasets. Finally, we also compare the proposed approach

This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

to a more traditional GMM-UBM (Gaussian Mixture Model-Universal Background Model) acoustic modeling technique that has been successful used in speaker and language identification problems.

## 2. Prof-Life-Log

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. The most popular use of the LENA device has been to capture the language environments of infants and young children, where the subject in question wears the unit. In our collection, the device is worn for the entire workday, and the audio data is captured continuously throughout the day. 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. For this study, we have annotated approximately 5 hours of data and used it for evaluation. This data was split into 10s segments and acoustic environment labels were assigned to each segment. Annotators were allowed to assign multiple labels to each segment. The annotation effort identified 50+ unique environment in total. 60% of the segments contained speech, while the other 40% contained background only (both single or mixed environments). In contrast, we have also used the LENA device to collect homogenous recordings of many different environments, such as, restaurant, walking (footsteps), large computing cluster, street, music, in-vehicle, office, *etc.* The homogenous recordings are different from the naturalistic recordings as they contain pure homogenous environments and none of the recordings contain speech. Similar to the naturalistic collection, the homogenous collection was also segmented into smaller audio chunks for the purpose of this study.

## 3. Proposed system

In this study, we propose an unsupervised technique towards audio environment detection. The proposed approach attempts to capture the acoustic signature of an acoustic environment. This process is shown in Fig.2. As shown in Fig. 2(a), a GMM (Gaussian Mixture Model) is trained using large quantities of diverse audio material. In principal, this process is similar to building a UBM (Universal Background Model) for speaker and language identification systems. In the proposed system, this GMM is used as the background acoustic model.

As shown in Fig. 2(b), the next step is to determine the acoustic signatures of audio environments. In what follows, we describe this process. Now, let $\vec{x}$ be the acoustic feature vector that is used to train the mentioned GMM. Also, it is assumed that the GMM consists of M-mixtures and $m_j$ is the $j^{th}$ mixture. Finally, assuming a generative model, let $P(\vec{x}|m_j)$ be the posterior probability of feature vector $\vec{x}$ being generated by mixture $m_j$. As mentioned in Sec. 2, the datasets are segmented into 10s segments for the purpose of indexing. Let the $k^{th}$ 10s segment be denoted by $V_k$. Assuming $V_k$ contains $N$ feature vectors, *i.e.*, $V_k = [\vec{x_1}\vec{x_2}\ldots\vec{x_N}]$, we compute the average posterior probability of mixture $m_j$ across all feature vectors in $V_k$ as,

$$q_m = \frac{1}{N}\Sigma_{i=1}^{N} P(\vec{x_i}|m_j). \quad (1)$$

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

$$Q = [q_1 q_2 \ldots q_M]^T. \quad (2)$$

The dimensions of $Q$ corresponding to the mixtures in the GMM that are more likely to generate the observed acoustic signal will contain higher values, and vice-versa. Hence, the vector $Q$ attempts to capture the unique acoustic signature of the signal. We term $Q$ as the acoustic signature vector (ASV). Fig. 3 shows example ASVs for restaurant and white noise, and compares it to speech. Following the mentioned procedure, the ASV is generated for all segments in the search dataset. Let the ASV for the $V_k$ ($k^{th}$ segment) be denoted by $Q_k$

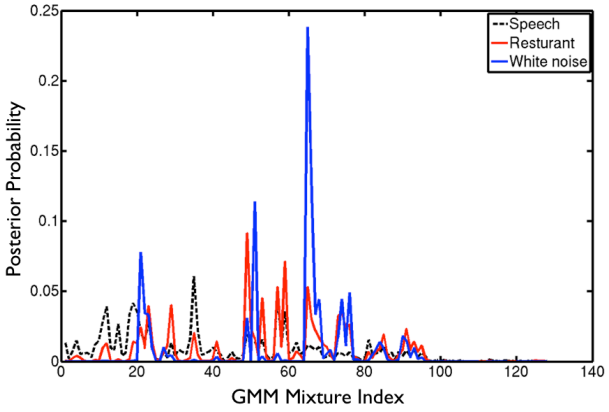


Figure 3: Example acoustic signature vectors (ASVs)

For searching, we follow a query by example process (see Fig. 2(c)). The user submits a segment and we find all the matching segments in our dataset. In order to find the matching segments, we first convert the user-submitted segment into its equivalent ASV following the procedure described. We denote the user-submitted ASV as $Q_{test}$. Now, the similarity between $Q_{test}$ and $Q_k$ can be computed by using the log-cosine distance measure, *i.e.*,

$$D(Q_{test}, Q_k) = \log\left(\frac{Q_k^T Q_{test}}{||Q_k||||Q_{test}||}\right). \quad (3)$$

where, $||.||$ is the norm operator. Lower values of $D(Q_{test}, Q_k)$ imply that $Q_{test}$ and $Q_k$ are similar and vice-versa. Now, the match *vs.* non-match binary decision can be made by comparing $D(Q_{test}, Q_k)$ to a threshold $\tau$.

## 4. Experiments

In this study, we have used MFCC (Mel-frequency cepstral coefficient) features to train the GMM. In particular, we used a frame duration of 32ms with 10ms skip, and 27-filterbanks. We used 12 static coefficients along with 12 velocity and acceleration coefficients to form a 36-dimensional feature vector.

For the homogenous dataset, we selected 5 different environments for experimentation, namely, computing cluster noise, restaurant, walking (footsteps), street, and music (in-vehicle audio system). Additionally, we also selected homogenous speech segments for comparison. We had 600 segments for each audio environment and speech, resulting in a total of 3600 homogenous segments. Using the mentioned GMM, we extracted the ASVs for these homogenous segments. In order to evaluate the proposed system, we followed the leave-one-out approach. A

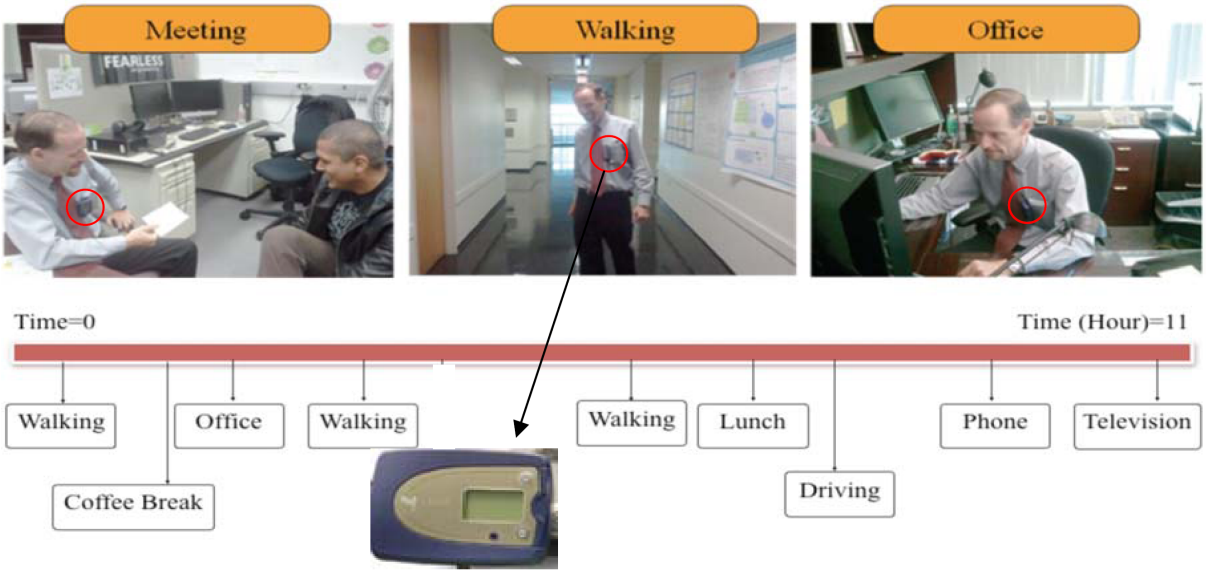


Figure 1: Data collection using the LENA unit: A single session consists of 10+ hours of audio recording with the speaker constantly carrying the unit. Speech is collected in a wide variety of backgrounds such as Cafeteria, Office, Meeting, Walking, Driving, *etc.* [13].

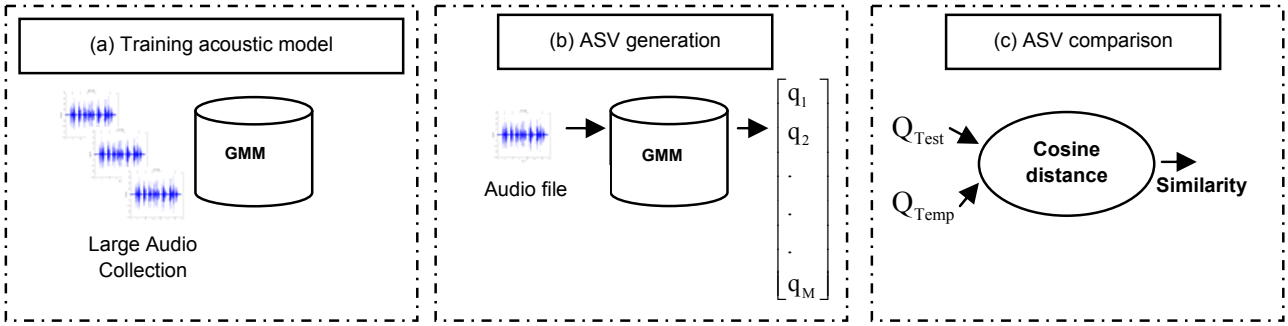


Figure 2: Proposed System: (a) A GMM is trained as a background acoustic model, (b) Acoustic Signature Vectors (ASVs) are extracted for the search dataset, and (c) Matching audio segments are extracted from the dataset based on a user-submitted query.

single segment was selected from the 3600 segments, and then compared to the other 3599 segments. This process was repeated for all 3600 segments.

For the naturalistic dataset, we chose 9 different environment for experimentation, namely, babble, clapping, indoors, laughing, typing, walking (footsteps), wind, restaurant, and outdoors. We also chose segments that contained speech and used it in our experiments. In total, we had 1800 segments and majority of the segments contained mixed environments (*i.e.*, the segment had more than one label). Similar to the homogenous dataset, the ASVs for the 1800 segments were extracted, and the system was evaluated using the leave-one-out approach.

We also trained a standard GMM-UBM system and compared its performance to the proposed system. The GMM-UBM system uses a standard MAP (maximum a-posteriori) approach that is common in speaker and language identification systems, and is briefly describe here. First, the background acoustic model is used as the UBM (Universal background model). Next, a segment is used to MAP-adapt the UBM to form the environment-model. Now, following the leave-one-out approach, a binary decision is made for the remaining segments, *i.e.*, the segments either belong to the UBM or the environment model. This is done by setting up a simple binary classification task where the likelihood of the UBM generating the segment is compared to the likelihood of environment model generating the segment.

Table 1: *EER% for our system vs. GMM-UBM baseline system on controlled set.*

| Mixtures | 8 | 32 | 128 |
|---|---|---|---|
| Proposed system | 14.32 | 10.27 | **7.32** |
| GMM-UBM | 17.06 | 15.18 | 16.38 |

## 5. Results and Discussion

Table 1 compares the performance of the proposed system with GMM-UBM for the homogenous dataset in terms of EER (equal error rate). It is seen that the the EER for the proposed system decreases with increasing number of GMM mixtures. In other words, a more complex acoustic model results is better

performance. In fact, the proposed system with a 128-mixture GMM gives the best result (7% EER) for the homogenous dataset. Additionally, it is also observed that the proposed system always outperforms the GMM-UBM system. Finally, unlike the proposed system, the GMM-UBM system performance is more or less constant with increasing model complexity. It is possible that short segments (of 10s length) are insufficient to adapt the UBM effectively. However, for the proposed system increasing model complexity may allow the acoustic signature vectors (ASVs) to be more distinct for different environments.

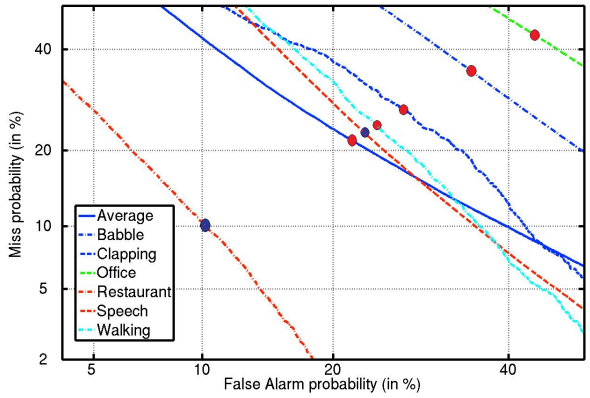


Figure 4: DET curve for proposed system for 36 dimensional MFCC features on uncontrolled data set.

Fig. 4 shows the DET (detection error tradeoff) curves for the proposed system when applied to the naturalistic dataset. The figure shows the performance for babble, clapping, office, restaurant, and walking environments. In addition, the DET curves for average performance and speech are also shown. It is observed that the best performance is obtained for the restaurant noise environment ($\sim$10% EER). On the other hand, the office environment has the worst performance ($\sim$45%). It is likely that the constituent acoustic events (such as music, silverware *etc.*) in the restaurant environment build a unique acoustic signature that is readily distinguished. In comparison, the office environment is dominated by silence (or quiet environment) which is probably not easily separated. Finally, the average EER for all environments is 22%. This result shows the inherent difficulty in separating real-world environments that tend to be mixed (unlike artificially collected homogenous environments).

Finally, Fig. 5 compares the performance of the proposed system to the GMM-UBM system for restaurant, walking, laughing and indoor environments. In addition, the average performance for all 9 environments is also shown. Similar to the homogenous dataset, once again the proposed system significantly outperforms the GMM-UBM system.

## 6. Conclusion

In this study, a novel environment detection algorithm was developed based on acoustic signature vectors (ASVs). The proposed system is an unsupervised technique where a user-submitted template can be matched to the search dataset using a query-by-example approach. In this manner, the system is useful even in previously unseen environments. The proposed algorithm was shown to outperform traditional GMM-UBM based system in an environment detection task using both homogenous and naturalistic datasets. In the future, we will work towards using more sophisticated acoustic models such as HMMs

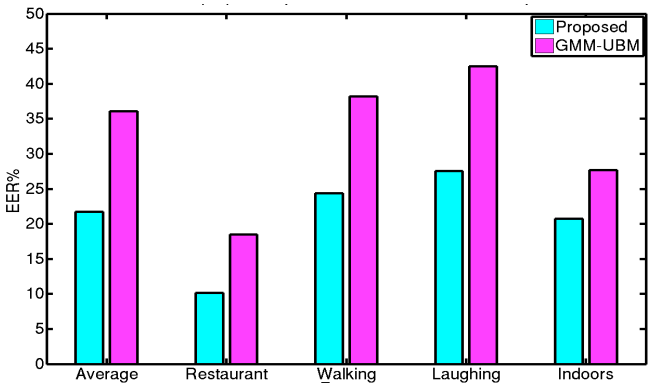


Figure 5: Comparing for proposed system to the baseline GMM-UBM system for 4 different environments on the naturalistic dataset. The average performance across all 9 environments is also shown.

(Hidden Markov Models) and employ techniques similar to keyword spotting for acoustic environment detection.

## 7. References


[1] Ogle, J.P. and Ellis, D.P.W., "Fingerprinting to identify repeated sound environments in long-duration personal audio recordings," ICASSP'07, pp. 233-236, 2007.

[2] Chu S., Narayanan S. and Kuo J. C-C"Environmental sound recognition with time-frequency audio features",IEEE Trans. Audio, Speech and Lang. Proc., 17(6):1142–1158, 2009.

[3] Akbacak M., Hansen, J.H.L"Environmental Sniffing: Noise Knowledge Estimation for Robust Speech System",IEEE Trans. Audio, Speech, and Lang. Proc., 15(2):465–477, 2007.

[4] Eronen A.J., Peltonen V.T., Tuomi J.T., Klapuri A.P, Fagerlund S., Sorsa T., Lorho G. and Huopaniemi J.,"Audio-based context recognition", IEEE Trans. Speech and Audio Proc., 14(1):321–329, 2006.

[5] Huang R. and Hansen J.H.L.,"Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora", IEEE Trans. Speech and Audio Proc., 14(3):907–919, 2006.

[6] Zhou B., Hansen, J.H.L"Efficient audio stream segmentation via the combined T static and Baysian information criterion",IEEE Trans. Speech and Audio Proc., 13(4):467–474, 2006.

[7] Slaney M.,"Semantic-Audio retrieval", International conference on acoustics, speech and signal proc. (ICASSP), Orlando, FL., USA, May, 13-17, 2002.

[8] Mesaros A., Heittola T., Eronen, A. and Virtanen T.,"Acoustic environment detection in real life recording", $18^{th}$ European Association for Signal Processing(EURASIP),1267-1271, 2010.

[9] Sundaram S. and Narayanam S.,"Audio retrieval by latent perceptual indexing", ICASSP, Las Vegas, NV., USA, March 30-April 4, 2008.

[10] Dong Zhao, Huadong Ma, Liang Liu, "environment classification for living environment surveillance using audio sensor networks," ICME, pp.528-533, 2010 IEEE International Conference on Multimedia and Expo, 2010

[11] Zhuang, X. and Zhou, X. and Hasegawa-Johnson, M.A. and Huang, T.S., "Real-world acoustic environment detection," Pattern Recognition Letters, Vol. 31, No. 12, pp. 1543-1551, 2010.

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

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