An i-vector Representation of Acoustic Environments for Audio-based Video Event Detection on User Generated Content

Benjamin Elizalde, Howard Lei, Gerald Friedland
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704, USA
benmael,hlei,fractor@icsi.berkeley.edu

***Abstract*—Audio-based video event detection (VED) on user-generated content (UGC) aims to find videos that show an observable event such as a wedding ceremony or birthday party rather than a sound, such as music, clapping or singing. The difficulty of video content analysis on UGC lies in the acoustic variability and lack of structure of the data. The UGC task has been explored mainly by computer vision, but can be benefited by the used of audio. The i-vector system is state-of-the-art in Speaker Verification, and is outperforming a conventional Gaussian Mixture Model (GMM)-based approach. The system compensates for undesired acoustic variability and extracts information from the acoustic environment, making it a meaningful choice for detection on UGC. This paper employs the i-vector-based system for audio-based VED on UGC and expands the understanding of the system on the task. It also includes a performance comparison with the conventional GMM-based and state-of-the-art Random Forest (RF)-based systems. The i-vector system aids audio-based event detection by addressing UGC audio characteristics. It outperforms the GMM-based system, and is competitive with the RF-based system in terms of the Missed Detection (MD) rate at 4% and 2.8% False Alarm (FA) rates, and complements the RF-based system by demonstrating slightly improvement in combination over the standalone systems.**

***Keywords*-Audio; i-vector; Video Event Detection; User Generated Content;**

I. Introduction

Social networks have transformed the internet into a sharing platform where users upload multimedia documents. Along with this change, recording gadgets have become widespread, and people have the ability to capture video in a greater capacity. This results in the amount of consumer-produced multimedia documents being increased rapidly on a minute-to-minute basis. However, all of these documents are of little value if they cannot be retrieved easily by consumers. For instance, a person who searches for video examples on how to give a marriage proposal may be at a loss if videos of marriage proposals can not be easily retrieved. Therefore, there is a need to have a VED system that can automatically analyze and fetch video content.

The VED task aims to identify videos with a semantically defined event, such as a marriage proposal. It is implicitly multimodal because events are characterized by audio-visual cues. Multimedia detection has been explored by computer vision using different features and techniques. However, audio has been under-explored, and the state-of-the-art audio-based techniques do not yet provide significant assistance to its video counterpart. Audio, however, can sometimes be more descriptive than video, especially when it comes to the descriptiveness of an event. For instance, the audio cue can quickly allow one to determine whether or not a marriage proposal was successful. Thus, there is great importance in exploring techniques to improve the use of audio for VED.

In the past, retrieval problems often suffered from limited training data. However, UGC videos can provide massive amounts of training data, because the videos are widely available, and contain metadata (i.e. such as the event described by the video, or the place in which the video was captured) which can provide ground truth labels. The audio of the UGC videos have the following characteristics – the presence of background noise, overlapped sounds, and diverse acoustic environments, among others.

This paper employs an i-vector based system for audio-based VED, as an attempt to address the challenges presented in UGC data. To understand better the system characteristics we include a performance comparison with conventional GMM-based and state-of-the-art RF-based systems, and indicate the benefits of the i-vector approach. The system provides new competitive results using audio features, and complements the RF-based system in combination. It also represents a simple, logical and scalable choice for the task, as it is a bag-of-frames (BOF) approach that does not rely on the use of acoustic concepts. The content of the paper is structured as follows. Section II presents the related work. Section III continues with the data. Section IV describes the i-vector-based system. Section V details the experimental setup. Section VI explains the results, and Section VII states the conclusion and future work.

II. Related work

There has been several past approaches to audio-based VED for UGC data. A technique closer to our work is an audio-based event detection paper based on a speaker verification system [1], which creates Gaussian Mixture

978-0-7695-5140-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ISM.2013.27

IEEE computer society

Models (GMM) for each event and classifies them using a likelihood ratio. Another example is a system [2] which extracts audio units automatically with a diarization system to create an audio word vocabulary, computes Term Frequency - Inverse Document Frequency (TF-IDF) histograms for each unit, and classifies them with a Support Vector Machine (SVM). A similar example is a system in [3] that creates an automatic audio word vocabulary with a RF algorithm, and computes TF-IDF histograms for each event based on the audio relevance. The histograms are then classified using a SVM. These two systems rely mainly on how distinctive the audio vocabulary represents each known event. These and other approaches to the task are inspired on successful speech-processing techniques.

The i-vector system has been successfully used also in tasks such as language recognition [4] and speaker diarization [5]on data captured in controlled environments. Although it has also been used in VED in [6] we considered relevant for the understanding of the i-vector performance and characteristics, to compare it with a conventional UBM-GMM system and a state-of-the-art technique.

III. Data

The data used in the experiments is the NIST TRECVID Multimedia Event Detection (MED) 2012 corpus, which contains consumer-produced video data. The entire corpus is comprised of a collection of 150,000 videos of about three minutes each. Some of the videos are used for acoustic event-class training, while the DEVT portion of the corpus is used for testing. For our experiments, only a subset of the corpus is used. The training set consists of 15 events from the Event Kits structure for a total of 2,024 video files, each containing audio. The test set consists of a total of 4,165 video files containing audio. From the test set 519 files belong to a specific event and the rest 3646 files do not belong to any of them. The video data may contain music, spontaneous speech, background noise, overlapped sounds, or other audio labeled as unintelligible in the annotations. Table I contains a summary of the event classes and the numbers of videos in each class.

Table I
SET OF EVENTS AND THE AMOUNT OF TRAINING AND TEST VIDEOS.

Code	Event	Train	Test
E001	Attempting a board trick	159	121
E002	Feeding an animal	160	119
E003	Landing a fish	119	83
E004	Wedding ceremony	125	86
E005	Working on a woodworking project	140	99
E006	Birthday party	173	2
E007	Changing a vehicle tire	109	1
E008	Flashmob gathering	173	0
E009	Getting a vehicle unstuck	129	0
E010	Grooming an animal	135	0
E011	Making a sandwich	123	0
E012	Parade	133	0
E013	Parkour	107	0
E014	Repairing an appliance	123	8
E015	Working on a sewing project	116	0

IV. The i-vector based event detection system

The i-vector system was initially developed by Dehak et al. [7], with an improvement made by Burget et al. [8]. It involves training a matrix T to model the total variability of a set of statistics for each audio track. The statistics primarily involve the first-order Baum-Welch (BW) statistics of the low-level acoustic feature frames (i.e. MFCCs) of each audio track. The BW statistics are in turn computed using a UBM. The Total Variability matrix T is low rank, and is used to obtain a low-dimensional vector characterizing the acoustic event of each audio track. Specifically, for each audio, the vector of first-order BW statistics M can be decomposed as follows, given the T matrix:

where m is the event-independent GMM, ω is a low - dimensional vector, referred to as the i-vector, and ϵ is the residual not captured by the terms m and $T\omega$. The i-vector can be thought of as a low-dimensional representation of the identity of each event class.

$$M = m + T\omega + \epsilon \tag{1}$$

For the TRECVID MED 2012 Event Kits training audio, one i-vector is obtained for each audio of each event class. For the test audio, one i-vector is obtained for each audio. The system then performs a Within-Class Covariance Normalization (WCCN) [9] on the i-vectors, which whitens the covariance of the i-vectors via a linear projection matrix. We followed an approach in [8], whereby a generative Probabilistic Linear Discriminant Analysis (pLDA) [10] log-likelihood ratio is used to obtain a similarity score between each test audio and each training event class, using the i-vectors. Because there are multiple audio samples per training event class, the i-vectors within each class are averaged such that each class is represented by one i-vector. The generative pLDA log-likelihood ratio for similarity score computation is shown below:

$$score(\omega_1, \omega_2) = logN\left(\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}; \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{bc}\\ \Sigma_{bc} & \Sigma_{tot}\end{bmatrix}\right) - logN\left(\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}; \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right)$$

where ω_1 and ω_2 are the two i-vectors, $N(\cdot)$ is the normal Gaussian probability density function, Σ_{tot} and Σ_{bc} are the total and between-class scatter matrices of the training i-vectors, prior to averaging. Hence, one score is obtained for each training event class versus test audio. The i-vector system involves several pre-trained components, such as the UBM, the T matrix, the WCCN projection matrix, and scatter matrices. All components were trained using the TRECVID MED 2012 Event Kits training audio. The Brno

University of Technology's JFA demo [11] and the ALIZE toolkit [12] are used to assist in system implementation.

The audio used is PCM-formatted, with a sample rate of 16kHz. The extracted acoustic features are the typical Mel-Frequency Cepstral Coefficients (MFCCs) C0-C19, with delta and double deltas, for a total of 60 dimensions. Each feature frame is computing using a 25 ms window, with 10 ms frame shifts. Short-time Gaussian feature warping using a three-second window is used, and temporal regions containing identical frames are removed.

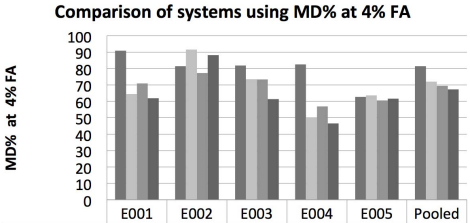
V. Experiments

A conventional audio-based event detection system [1], based on the GMM-UBM approach, is used to provide a baseline comparison with our results. The system has two steps – the creation of the event models (training), and the scoring (testing). In the first step, the system receives the audio representing a known event as input, and extracts the MFCC acoustic features. It then performs maximum a posteriori (MAP) adaptation to adapt a pre-trained event-independent 256-mixture GMM, known as the Universal Background Model (UBM), to create a 256-mixture GMM for each event. In the second step, a log-likelihood ratio is used to obtain a similarity score between each event-dependent GMM, and the acoustic features of each test audio. The UBM is used in the likelihood-ratio computation for score normalization. Note that the UBM is trained using the audio of the TRECVID MED 2012 Event Kits.

A state-of-the-art audio-based system [3], based on the use of RF decision trees, is also included in the results comparison. First, the system extracts MFCC features and learns a dictionary based on the RF outputs, using the training data. The output of each leaf node in the RF is used as an audio word. Second, the audio words are weighted according to the logarithm of the TF-IDF of the audio words. Lastly, histograms of the weighted audio words are used to represent each audio, and classified using an SVM.

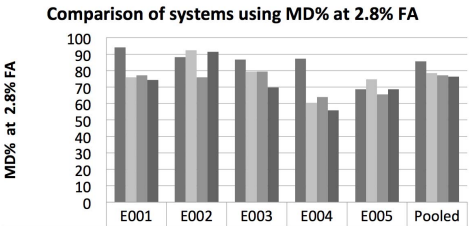
Lastly, the score-level combination performance of the i-vector and RF-based systems are included in the comparison. The scores of the standalone systems are first scaled to fall within the range of 0 to 1. Combination is performed by simply adding the standalone scores.

The 15 EventKits training data and the DEVT test data, as described in Section III, were used. There are 256-mixture UBMs used for the i-vector and GMM-UBM systems, and 400-dimensional i-vectors are used for the i-vector system. The evaluation metrics utilized to compare the systems are the Missed Detection (MD) rates at 4% and 2.8% False Alarm (FA) rates. These FA rates are based on the Pre-Specified Events metrics from the TRECVID (MED) evaluations [13], of years 2012 and 2013. The MD is the percentage of matched-event scores that are mis-classified as non-matched event scores, when the percent of non-matched event scores being mis-classified as matched-event scores is set at 4% or 2.8%, using a scoring threshold.



	E001	E002	E003	E004	E005	Pooled
GMM-UBM (Left)	90.9	81.5	81.9	82.6	62.7	81.5
Rnd-Forest	64.5	91.6	73.5	50	63.7	72
i-vector	71	77.3	73.4	56.9	60.6	69.5
RF + i-vector (Right)	61.9	88.2	61.4	46.5	61.6	67.3

Figure 1. The MD at 4% FA performances of the GMM-UBM, RF, i-vector and combined RF and i-vector systems, for five individual event categories and the pooled category (E001-E005).



	E001	E002	E003	E004	E005	Pooled
GMM-UBM (Left)	94.2	88.2	86.7	87.2	68.7	85.6
Rnd-Forest	76	92.4	79.5	60.5	74.7	78.5
i-vector	77.1	76	79.5	63.9	65.6	77.2
RF + i-vector (Right)	74.3	91.5	69.8	55.8	68.6	76.3

Figure 2. The MD at 2.8% FA performances of the GMM-UBM, RF, i-vector and combined RF and i-vector systems, for five individual event categories and the pooled category (E001-E005).

VI. Results

The MD results for the detection performances of the individual events in the training data, along with a pooled set of events, are shown in Figures 1 and 2. Figure 1 shows the MD performance at a 4% FA, and Figure 2 shows the MD performance at 2.8% FA. Only the first five events are evaluated and included in the figures, because events 6 to 15 lacked sufficient test data.

In Figure 1, the i-vector system outperforms the GMM-UBM system in each individual event category and the pooled category (E001-E005), achieving a 14.7% relative MD improvement at 4% FA (69.5% vs. 81.5%) for the pooled category. The i-vector system also slightly outperforms the RF-based system for the pooled category, achieving a 3.7% relative MD improvement (69.2% vs. 72%). The combination of the i-vector and RF-based systems performs roughly similarly compared to the i-vector system standalone (67.3% vs. 69.5%), and gives a 6.5% relative improvement over the RF-based system (67.3% vs. 72%).

According to Figure 2, the i-vector system also outperforms the GMM-UBM system in each individual event category and the pooled category, achieving a 9.8% relative MD improvement at 2.8% FA (77.2% vs. 85.6%) for the pooled category. The i-vector system performs about the same as the RF-based system for the pooled category for MD at 2.8% FA (77.2% vs. 78.5%). For this metric, the combination of the i-vector and RF-based systems does not significantly improve over the standalone systems (76.3% for the combined system, vs. 78.5% for the RF-based system).

One reason the i-vector system is perhaps able to improve results is that it can capture the acoustic event characteristics contained in the audio using a low-dimensional vector (see Section IV). Furthermore, the WCCN and pLDA system components normalize for the within- and between-class i-vector scatter of the events, which accounts for cases when the same event contains distinctive audio across videos, and when different events contain similar audio.

The i-vector system is also fairly efficient in terms of computation time, which is crucial for large-scale data tasks. On an Intel-Xeon-E5-2660, 64-bit 2.2 GHz processor, the system took a total of 6.1 hours for i-vector extraction and similarity score generation. Given the 62,475 total similarity scores that are generated, each score took 0.35 seconds to compute. Aside from i-vector extraction and scoring, the system also used 1.6 hours to generate the UBM, and 0.14 hours to generate the T-matrix. Both the UBM and T-matrix are event-independent, and can be used repeatedly for successive runs of the system.

VII. Conclusion

This works shows that the i-vector system is a competitive approach for audio-based VED on user-generated video content. The results reveal significant improvements in comparison to the conventional GMM-based system, and competitive performance in comparison to the RF-based system. Furthermore, the i-vector system is able to complement the RF-based system in combination. The strength of the algorithm is that it takes into account the within- and between-event acoustic scatter using WCCN and pLDA, allowing the algorithm to account for scenarios where multiple videos of the same event have different acoustic characteristics, and where videos from different events have similar acoustic characteristics. Therefore, the technique provides a valid approach not only for tackling the event detection task itself, but also for handling the difficulties of UGC data.

Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government. Thanks to Adam Janin, Korbinian Riedhammer, Nils Peter and Jaeyoung Choi for their advice.

References

[1] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakaran, "Acoustic Super Models for Large Scale Video Event Detection," in *ACM Multimedia*, 2011.

[2] B. Elizalde, G. Friedland, H. Lei, and A. Divakaran, "There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[3] P.-S. Huang, R. Mertens, A. Divakaran, G. Friedland, and M. Hasegawa-Johnson, "How to put it into words - Using random forests to extract symbol level descriptions from audio content for concept detection," in *ICASSP*, 2012.

[4] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *in Proceedings of Interspeech*, 2011.

[5] D. T. T. Javier Franco-Pedroso, Ignacio Lopez-Moreno and J. Gonzalez-Rodriguez, "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation," in *in FALA "VI Jornadas en Tecnologia del Habla" and II Iberian SLTech Workshop*, 2010.

[6] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Compact audio representation for event detection in consumer media," in *INTERSPEECH*, 2012.

[7] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, UK, 2009.

[8] L. Burget, P. Oldřich, C. Sandro, O. G., P. M., and N. Brümmer, "Discriminantly trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings of ICASSP*, Brno, Czech Republic, 2011.

[9] A. O. Hatch, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *Proceedings of ICASSP*, Toulouse, France, 2006.

[10] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of ECCV*, 2006, pp. 531–542.

[11] O. Glembek, "Joint factor analysis matlab demo," http://speech.fit.vutbr.cz.

[12] J. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proceedings of ICASSP*, 2005.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Queenot, "Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*. NIST, 2012.