AN I-VECTOR BASED APPROACH FOR AUDIO SCENE DETECTION

Benjamin Elizalde, Howard Lei, Gerald Friedland

International Computer Science Institute
1947 Center Street
94704 Berkeley, CA, USA
{benmael|hlei|fractor}@icsi.berkeley.edu

Nils Peters

Qualcomm Technologies Inc.
5775 Morehouse Drive
92121 San Diego, CA, USA
npeters@qti.qualcomm.com

ABSTRACT

The IEEE-ASSP Scene Classification challenge on user-generated content (UGC) aims to classify an audio recording that belongs to a specific scene such as busystreet, office or supermarket. The difficulty of scene content analysis on UGC lies in the lack of structure and acoustic variability of the data. The i-vector system is state-of-the-art in Speaker Verification and Scene Detection, and is outperforming conventional Gaussian Mixture Model (GMM)-based approaches. 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 reports our results in the challenge by using a hand-tuned i-vector system and MFCC features. Compared to the MFCC+GMM baseline system, our system increased the classification accuracy by 26.4% to about 65.8%. We discuss our approach and highlight parameters in our system that showed to significantly improved our classification accuracy.

Index Terms— User Generated Content, Scene Detection, Event Detection, Audio, GMM, i-vector, MFCC

1. INTRODUCTION

The Audio and Multimedia Research Group of the International Computer Science Institute in Berkeley develops and refines computational algorithms, systems, and methods to handle large amount of UGC composed of multiple types of data and meta-data, such as videos, social media feeds, and geo-tags. We are actively involved in content analysis projects such as Mediaeval [1], TRECVID Multimedia Event (Scene) Detection (MED) [2], acoustic segmentation [3], and room identification [4].

Scene detection has been explored by computer vision using different features and techniques. However, audio has been underexplored, 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.

In the past, retrieval problems often suffered from limited training data. In contrast, UGC (such as you-tube videos) is generally available in large scale for content analysis. UGC is also known to be unstructured, in the sense of low-quality recordings, background- and environmental noise, and variation of the acoustic content itself.

The dataset provided with this IEEE-AASP Classification challenge gives us an opportunity to gain experience with a rather small UGC dataset. It comprises 100 files, ten 30 seconds audio files for each of the ten scenes. The audio files are binaurally captured, incorporating the binaural cues of three unknown heads at 44.1 kHz and 16-bit PCM.

2. OUR SCENE DETECTION SYSTEM

Our challenge contribution employs an i-vector based system for the audio scene detection. This approach has been previously tested with MED data [5]. It outperformed the GMM-based system, and was competitive with the Random Forest-based system in terms of the Missed Detection rate at 4% and False Alarm rates at 2.8%.

2.1. The i-vector approach

The i-vector system was initially developed by Dehak et al. [6], with an improvement made by Burget et al. [7]. The system 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 statistics of the low-level acoustic feature frames (i.e., MFCCs) of each audio track. The Baum-Welch 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 Baum-Welch statistics M can be decomposed as follows, given the T matrix:

$$M = m + T\omega + \epsilon \quad (1)$$

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.

For the Challenge, five stratified folds were created, with 80 audio files for training and 20 audio files for testing. One i-vector is obtained for each audio file. The system performs a Within-Class Covariance Normalization (WCCN) [8] on the i-vectors, which whitens the covariance of the i-vectors via a linear projection matrix. We followed an approach in [7], whereby a generative Probabilistic Linear Discriminant Analysis (pLDA) [9] 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 using the above approach. The i-vector system involves several pre-trained components, such as the UBM, the T matrix, the WCCN projection matrix, and the total- and between-class scatter matrices. All such components were trained using the fold's corresponding training audio. The Brno University of Technology's JFA Matlab demo [10] is used to assist in the i-vector system development. The open-source ALIZE toolkit [11] is used to train the UBM.

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, extracted using the HTK tool [12]. Each feature frame is computing using a 25 ms window, with 10 ms frame shifts. A frequency range of 60-20000 Hz and 52 triangle filter-banks were selected. Short-time Gaussian feature warping using a three-second window [13] is used, and temporal regions containing identical frames are removed.

2.2. Audio pre-processing

To take advantage of additional cues embedded in the binaural recordings of this dataset, while maintaining the system's one-channel processing architecture, we extract from each audio files four different monaural versions and concatenate them, resulting in a one-channel file with a duration of two minutes. The four different monaural versions are:

1. Left channel
2. Right channel
3. Channel difference: left channel - right channel
4. Channel average: (left channel + right channel)/2

We hope that the MFCC features extracted from these concatenated versions will provide more useful cues to the i-vector system, compared to extracting MFCCs from just one channel.

3. RESULTS

Our final system achieved an accuracy of $65.8\% \pm 4.8\%$ (95% C.I.) averaged across four different 5-fold stratified cross-validations using the provided training dataset. The results suggests that we could increased the baseline accuracy by 26.4% compared to the baseline system.

The confusion matrices from the baseline system [14] and from our i-vector system are shown in Table 1. We choose to present the confusion matrix from the least accurate of our four 5-fold stratified CVs. A classification accuracy of 80% or better was achieved for the scenes *busystreet*, *openairmarket*, and *park*. The least accurate results are related to the scenes *tube* (40%), and *tubestation* (30%). Compared to the baseline results our system has a similar or higher accuracy in six classes. Especially for *park*, *busystreet* and *restaurant*, our system achieved a higher classification score. While the supermarket scene is the true scene most often misclassified in the baseline system (83.3%), our system performs reasonably well (37.5%).

bus	**9**	-	-	-	-	-	-	-	1	-
busystreet	-	5	-	2	-	-	1	-	-	2
office	-	-	**8**	-	1	-	-	1	-	-
openairmarket	-	-	-	8	-	-	1	1	-	-
park	-	-	2	1	3	3	-	1	-	-
quietstreet	-	-	-	2	2	4	-	2	-	-
restaurant	-	-	-	2	-	-	3	3	-	2
supermarket	1	-	1	2	1	-	1	2	-	2
tube	-	-	-	-	-	-	2	-	**6**	2
tubestation	-	-	-	2	-	-	-	2	2	**4**

baseline ↑ / i-vector ↓	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation
bus	6	-	1	2	-	-	1	-	-	-
busystreet	-	**9**	-	-	-	-	-	-	-	1
office	-	-	6	-	2	1	-	-	1	-
openairmarket	1	-	-	8	-	-	1	-	-	-
park	-	-	2	-	**8**	-	-	-	-	-
quietstreet	-	2	1	2	-	**5**	-	-	-	-
restaurant	-	-	-	1	-	-	**7**	2	-	-
supermarket	1	1	1	1	-	-	-	**5**	-	1
tube	-	1	-	1	1	-	-	-	4	3
tubestation	-	2	-	1	-	1	-	1	2	3

Table 1: Confusion matrices for baseline system (top) and our i-vector system (bottom). Rows are ground-truth labels. **In bold:** the system with the higher classification score.

4. DISCUSSION

When separating the ten scenes into *indoor* and *outdoor*[1] categories and comparing the achieved accuracy those two categories, it becomes clear that our system outperforms the baseline system for outdoor scene classification. However, it also shows that our system has difficulties with the six indoor scenes, see Figure 4. Moreover, as observable in the confusion matrix (Table 1) the indoor recordings often gets mislabeled as the outdoor scenes busystreet or openairmarket. This indoor-outdoor confusion must be prevented to increase our system's accuracy, maybe by employing additional binaural features, such as those based on the interaural cross correlation (IACC).

[1]Outdoor scenes: busystreet, openairmarket, park, quietstreet; Indoor scenes: the remaining six scenes.

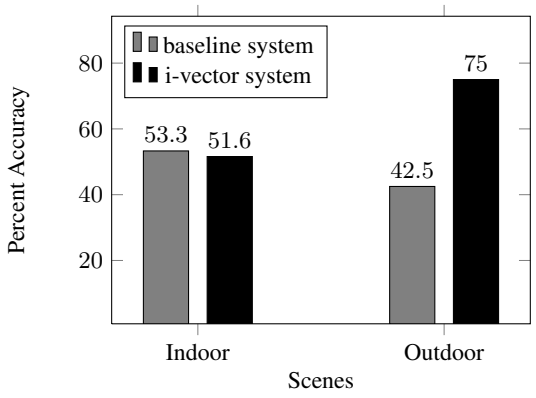

Figure 1: Mean accuacy for indoor and outdoor scenes

5. CONCLUSION AND FUTURE WORK

These results show the feasibility of using an MFCC+i-vector system for a scene classification task and significant improvements in comparison to the conventional GMM-based system. The classification accuracy of 80% or better was achieved for the scenes busystreet, openairmarket, and park, while the scene tubestation received the least classification accuracy (30%).

To potentially improve the accuracy for those classes, additional features such as those based on the modulation spectrogram might be beneficial and could be added to the system. Furthermore, implicit binaural features, such as the interaural cross correlation coefficient (IACC) could help to improve the differentiation of indoor/outdoor characteristics.

The i-vector system provides a valid approach not only for tackling the scene detection task itself, but also for handling the difficulties of UGC data.

6. ACKNOWLEDGMENT

Thanks to Brno University of Technology for providing the Joint Factor Analysis matlab scripts used in our i-vector system.

7. REFERENCES

[1] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran, "Multimodal location estimation of consumer media: Dealing with sparse training data," in *Proceedings of the IEEE ICME*, 2012, pp. 43–48.

[2] 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." NIST, USA, 2011.

[3] B. Elizalde and G. Friedland, "Lost in Segmentation: Three Approaches for Speech/Non-Speech Detection in Consumer-Produced Videos ," in *Proceedings of the IEEE ICME*, 2013.

[4] N. Peters, H. Lei, and G. Friedand, "Name That Room: Room Identification Using Acoustic Features in a Recording," in *Proc. of ACM Multimedia*, Nara, Japan, 2012.

[5] B. Elizalde, H. Lei, Q. Yu, A. Divakaran, and G. Friedland, "An i-vector based approach for improved video event detection using audio," *submitted*, 2013.

[6] 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*, Brighton, UK, 2009.

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

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

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

[10] O. Glembek, "Joint factor analysis matlab demo," http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo/.

[11] J. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proceedings of ICASSP*, vol. 1, 2005, pp. 737–740.

[12] S. Young and S. Young, "The htk hidden markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of Speaker Odyssey*, Crete, Greece, 2001.

[14] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "IEEE AASP challenge: Detection and classification of acoustic scenes and events," 2013.