CAPTURING THE ACOUSTIC SCENE CHARACTERISTICS 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

Scene detection on user-generated content (UGC) aims to classify an audio recording that belongs to a specific scene such as busy street, office or supermarket rather than a sound such as car noise, computer keyboard or cash machine. The difficulty of scene content analysis on UGC lies in the lack of structure and acoustic variability of the audio. 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 on the IEEE-AASP Scene Classification Challenge dataset. Compared to the MFCC+GMM baseline system, our approach increased the classification accuracy by 26.4% relative, to 65.8%. We discuss our approach and highlight parameters in our system that significantly improved our classification accuracy.

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

1. INTRODUCTION

Scene detection aims to identify recordings with a semantically defined scene. This task 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. For instance, the audio cue can quickly allow one to determine whether a street is busy or quite. Thus, there is great importance in exploring techniques to improve the use of audio for scene detection.

In the past, retrieval problems often suffered from limited training data. In contrast, UGC (such as YouTube videos) is generally available in large scale for content analysis. UGC is also known to be unstructured, in the sense that it contains low-quality recordings, background- and environmental acoustics, and overlapping and variable sound durations. The dataset provided by the IEEE-AASP Classification Challenge [1] gives us an opportunity to gain experience with a rather small UGC dataset.

This paper employs an i-vector based system for audio-based video event detection, as an attempt to address the challenges presented in UGC data. The system provides competitive results using audio features, for the small UGC dataset. The approach 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 audio concepts or sounds. A description of the hand-tuned system and the results is included.

This paper is structured as follows: Section 2 presents the related work. Section 3 continues with the data. Section 4 describes the scene detection system. Section 5 details the experimental setup. Section 6 presents and explains the results, and Section 7 states the conclusion and future work.

2. RELATED WORK

There have been different approaches to audio-based scene detection for UGC audio. Some of them were designed towards the NIST evaluation called TRECVID Multimedia Event (Scene) Detection (MED) [2]. An early system [3] creates Gaussian Mixture Models (GMM) for each scene and classifies them using a likelihood ratio. A more recent example is a system [4] that 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 [5] 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. The paper from [6] employs an i-vector technique combined with cosine kernels or chi-square kernels in an SVM. Our audio-based scene detection system based on the use of i-vectors has been previously tested with MED data [7].

In speaker verification, the i-vector system [8, 9] is now the state-of-the-art. The approach used in the system conveys the audio class characteristics, among other information, such as transmission channel, acoustic environment, or acoustic content. These properties make the i-vector system a meaningful approach for a detection task on UGC data. The technique has been successfully used also in tasks such as language recognition [10] and speaker diarization [11] on data captured in controlled environments. However, the i-vector system has not been thoroughly explored on audio-based scene detection using UGC audio.

3. DATA

The IEEE-AASP Classification Challenge provided the audio set used in the experiments. The set comprises 100 audio files, ten 30 second audio files for each of the ten scenes. The ten scenes are shown in Table 1. The audio may contain spontaneous speech, background noise, overlapped sounds, or other unintelligible audio. The set provides no extra annotations other than the scene name of each sample. The audio files are binaurally captured, incorporating the binaural cues of three unknown heads at 44.1 kHz and 16-bit PCM.

4. AUDIO-BASED SCENE DETECTION SYSTEM

The i-vector system was initially developed by Dehak et al. [8], with an improvement made by Burget et al. [9]. 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:

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

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) [12] on the i-vectors, which whitens the covariance of the i-vectors via a linear projection matrix. We followed an approach in [9], whereby a generative Probabilistic Linear Discriminant Analysis (pLDA) [13] 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:

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

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 [14] is used to assist in the i-vector system development. The open-source ALIZE toolkit [15] 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 [16]. 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 [17] is used, and temporal regions containing identical frames are removed.

5. EXPERIMENTS

The experiments consisted of computing an average across four 5-fold stratified cross-validations using the provided 100 file audio set. Each fold consisted of a randomly selected and unique subset of eight files for training, and two for testing, for each scene class. The reason for performing several cross-validations was to provide greater statistical significance in the results for the small audio set. The baseline results presented here were taken from the IEEE-AASP Classification Challenge [1], which employed a conventional UBM-GMM and likelihood ratio system to process a 5-fold stratified cross validation set. The metrics used to compare performance are classification accuracy, and the corresponding confusion matrices for further analysis.

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 file four different monaural versions and concatenated 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

The extracted MFCCs from these concatenated versions are expected to provide more useful cues for the i-vector system, compared to extracting MFCCs from just one channel.

6. RESULTS AND DISCUSSION

The final system achieved an accuracy of $65.8\% \pm 4.8\%$ with 95% confidence interval (C.I.) averaged across four different 5-fold stratified cross-validations. The results suggest an increased accuracy of 26.4% relative compared to the baseline system with $52\% \pm 13\%$ with 95% C.I.

The confusion matrices from the baseline system [18] and from our i-vector system are shown in Table 1. For an understated comparison with the baseline system, we deliberately show the confusion matrix from the least accurate of our four 5-fold stratified cross validation sets.

A classification accuracy of 80% or better was achieved for the scenes *busystreet*, *openairmarket*, and *park*. The least accurate numbers are related to the scenes *tube* (40%), and *tubestation* (30%). Compared to the baseline results, our system has a similar or higher accuracy in 6 of the 10 classes. Our system performed especially well for classes such as *park*, *busystreet* and *restaurant*.

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.

Table 2: There are four outdoor and six indoor scenes.

Outdoor	**Indoor**
busystreet	bus
openairmarket	office
park	restaurant
quietstreet	supermarket
	tube
	tubestation

While the supermarket scene is the scene that is most often misclassified in the baseline system (83.3% false positives), our system performs reasonably well (only 37.5% false positives).

When separating the ten scenes into *indoor* and *outdoor* categories as shown in Table 2 and comparing the achieved accuracy of the two categories, it becomes clear that our system outperforms the baseline system for outdoor scene classification. However, our system has difficulties with the six indoor scenes, as shown in Figure 1. Moreover, as observable in the confusion matrix in 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, for instance by employing a room identification system [19], or by using additional binaural features, such as those based on the inter-aural cross correlation (IACC). The approach could also be used in tandem with other audio-based techniques that addresses the i-vector weaknesses to achieve better results.

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 i-vector, described in Section 4. Furthermore, the WCCN and pLDA components of the the system normalize for the within- and between-class i-vector scatter of the events, which accounts for cases when the same scene contains distinctly different audio in different recordings, and when different scenes contain similar audio.

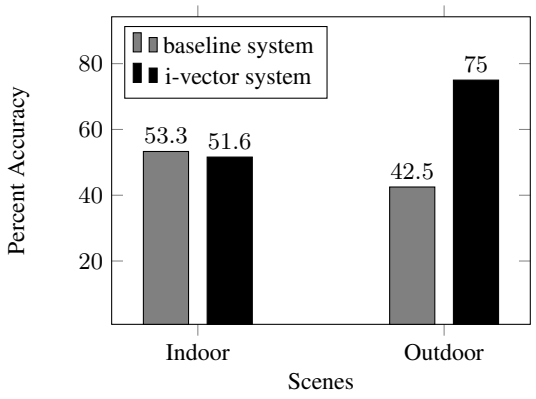

Figure 1: The mean accuracy for indoor and outdoor scenes show that our i-vector system outperforms the baseline system for outdoor scene classification.

7. CONCLUSION AND FUTURE WORK

Scene detection on user-generated content (UGC) aims to classify an audio recording that belongs to a specific scene such as busystreet, office or supermarket, rather than a sound such as car noise, computer keyboard or cash machine. The difficulty of scene content analysis on UGC lies in the lack of structure and acoustic variability of the audio. 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.

The results on this paper show the feasibility of using an MFCC+i-vector system for a scene detection task and the 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 of those classes, additional features such as those based on the modulation spectrogram could be beneficial and could be added to the system. Furthermore, implicit binaural features, such as the inter-aural 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.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] D. Giannoulis, E. Benetos, D. Stowell, and M. D. Plumbley, "IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Public Dataset for Scene Classification Task," 2012, queen Mary University of London.

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

[4] 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.

[5] 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.

[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*. ISCA, 2012. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#ZhuangTWNPN12

[7] 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.

[8] 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.

[9] 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.

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

[11] 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, pp. 415—418.

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

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

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

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

[16] 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.

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

[18] 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.

[19] 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.