Segmental-GMM Approach based on Acoustic Concept Segmentation

Diego Castan[1], *Murat Akbacak*[2]

[1]University of Zaragoza, Spain
[2]Microsoft, Sunnyvale, CA, USA

dcastan@unizar.es, murat.akbacak@ieee.org

Abstract

The amount of multimedia content is increasing day by day, and there is a need to have automatic retrieval systems with high accuracy. In addition, there is a demand for event detectors that go beyond the simple finding of objects but rather detect more abstract concepts, such as "woodworking" or a "board trick." This article presents a novelty approach for event classification that enables searching by audio concepts from the analysis of the audio track. This approach deals with the *acoustic concepts recognition* (ACR) creating a trained segmentation instead a fixed segmentation as *segmental-GMM* approach with broad concepts. Proposed approach has been evaluated on NIST 2011 TRECVID MED development set, which consists of user-generated videos from the Internet, and has shown a EER of 40%.

Index Terms: Multimedia event detection(MED), acoustic concept recognition, segmental-GMM

1. Introduction

In recent years, there has been a growing demand for high-accuracy multimedia retrieval systems due to the popularity of the video-sharing websites. For a multimedia retrieval task, video features can determine the general content of a video. However, the audio track of the video can also be critical. Consider the case of a tennis match video where a special event, like a new point, may occur. Audio analysis provide a complementary information to detect this specific event (detecting applause or cheering) that would be significantly more difficult to detect with image/video analysis. The Text Retrieval Conferences Video Retrieval Evaluation (TRECVID) addresses the problem of *Multimedia Event Detection* (MED) requiring a system that can search user-submitted quality videos for specific events [1].

Different applications for acoustic processing on multimedia videos have recently been described in the literature. These applications have been used as acoustic concept detectors in different scenarios. In [2] and [3], authors developed an SVM-based system and an HMM-based system, respectively, to classify different acoustic sounds (e.g., steps, door slams, or paper noise) in the meeting room environment. Both approaches use the CHIL-2007 database in which the acoustic concepts are isolated and recorded in a controlled environment [4].In the multimedia content analysis domain, most of the studies are concentrated on finding small events or objects rather than entire concepts. Very good summaries are provided by [5] and [6]. However, spoken concepts approaches are commonly used to detect multimedia events like [7] and [8].

Audio concept extraction approaches explored under different multimedia retrieval and content analysis projects can be grouped into two categories: (1) unsupervised and (2) supervised approaches from the perspective of modeling acoustic concepts. In the first group, one popular unsupervised approach is the Bag-of-Audio-Words (BoAW) method. In this approach, all frame-level features are clustered via vector quantization (VQ), and then VQ indices are used as features within a classifier to model audio content ([9, 10]). Other unsupervised approaches are focused on segmenting the audio track, and clustering the segments to form atomic sound units and then word-like units [11, 12], or modeling the segments with i-vectors [13] or GMM super-vectors [14] which are methods borrowed from speaker identification. In the second group of approaches, audio concept/event models are trained using annotated data [15, 16]. For example, in [15], fixed-duration segments are represented with segmental-GMM vectors where each element in the vector is a GMM score calculated from a pretrained GMM that corresponds to an annotated concept label. In [16], authors model acoustic concepts by training SVMs on 10sec audio segments which are annotated with generic concept labels (e.g., indoor vs. outdoor), and they use detected acoustic concept labels as features for multimedia event detection task. Some systems employ a combination of different approaches like in [17] where authors combine automatic speech recognition with broad-class acoustic concepts. Although the first group of approaches has the advantage of not requiring labeled acoustic event/concept data, these approaches do not present semantic labels to allow semantic searches. This is an important aspect for tasks such as multimedia event detection when the number of examples for multimedia event types becomes quite small. Therefore supervised acoustic concept detectors will be useful to tackle this problem.

This paper presents a specific study with two approaches to model five broad acoustic concepts as a MED features: segmental-GMM vectors [15] as a baseline, and a set of features based on *Acoustic Concepts Recognition* with HMMs. The broad acoustic concepts were chosen to describe sounds of different nature (people sounds, machine noises ...) and be able to model general concepts to provide a tool for retrieval information with no prior knowledge of specific acoustic events. The first part of this paper shows the classification accuracy over the isolated broad concepts. Secondly, an experiment with two extra concepts (music and speech) indicates the difficulty to provide the segmentation of a user video in general concepts. Finally, we employ an HMM-based *acoustic concept recognition* (ACR) system to segment the audio signal. The segmental information is used as features in SVM-based classification for multimedia event detection (MED) task. This approach is different from the previously mentioned supervised techniques [15, 16] in several ways. First, we do not use any fixed segmentation, but instead use recognition to extract acoustic concept segments dynamically. The second difference is that the models are not trained with specific acoustic concepts that may produce a system very constrained for a specific task.

Abbr.	Full Name	# Train	# Test
E001	Attempting a board trick	91	32
E002	Feeding an animal	81	30
E003	Landing a fish	69	26
E004	Wedding ceremony	66	25
E005	Woodworking project	77	25

Table 1: Video event class abbreviations (Abbr.) and full names along with the number of positive samples appearing in the training and test sets

The remainder of this paper is organized as follows: the TRECVid2011 dataset and the acoustic concepts annotations are described next. Section 3 deals with the audio features and the acoustic concepts classification and recognition (segmentation and classification). The baseline of MED task using segmental-GMM vectors and the ACR system are provided in Section 4. Finally, conclusions will be presented in Section 5.

2. Data set and Annotations

2.1. TRECVid 2011

The Text Retrieval Conferences Video Retrieval Evaluation (TRECVID) [1] focuses on the problem of Multimedia Event Detection (MED) in website quality videos for hard-to-detect events (e.g., Landing a fish). The evaluation dataset consists of non-professional videos collected from the internet with high variability and short duration (a couple of minutes). Fifteen different video event categories can be found in the database with only five of those categories available for testing purposes in this study.

To develop and evaluate our proposed approach, we use three sets of data: first set (*train-1*) is for training the acoustic concept models, second set (*train-2*) is for training the MED classifiers after extracting acoustic concept indexes on this data and using them as MED features, and the third set (*test*) is for testing the system. These sets are the same used in [15] and [9] to be able to provide fair comparison to previously published works. There is a total of 2640 videos in the test set and 7881 in the training set. Table 1 shows, for each of the five video events, the numbers of positive samples in the test and training sets. Note that the categories grouped several videos. For example "feeding an animal" includes animals from different species and , therefore, different animal sounds.

2.2. Acoustic Concepts Annotations

Because the ultimate goal of the system is to perform detection of multimedia events on the videos using the recognition of acoustic concepts, it has been created an initial set of labels of acoustic concepts to be useful in discriminating the five video event classes presented in Table 1 while also being clear and understandable for the annotators.

Broad Acoustic Concepts	Abbr.
1. Crowds and audience	(CA)
2. Animal sounds	(AN)
3. Repetitive sounds	(RS)
4. Machine noise	(MN)
5. Environmental sound	(ES)
6. Music	(MU)
7. Speech	(SP)

Table 2: Broad acoustic concepts and abbreviations

The acoustic concepts are divided in five broad classes as Table 2 shows. These broad classes have been extended with Speech and Music classes because most of the videos contain speech or music as the predominant audio. In fact, some of the five acoustic concepts are overlapped with speech or music barely audible in the background. However, those segments were annotated as that acoustic concept. The following section presents the results on the classification, and recognition of the broad classes, showing how difficult is to create a well-trained model for these acoustic concepts due to the high variance of the audio.

3. Acoustic Concepts Recognition

To model the acoustic concepts, we used a HMM/GMM-based system. As it was described on the last section, to train and test these models, a subset of the National Institute of Standards and Technology (NIST) is provided for the TRECVID evaluation 2011. This set is composed of 1536 videos (47 hours approximately) averaging 1.8 minutes per file. This section is organized follows: we describe the front-end audio features used in this approach and the acoustic concepts to train the models. Also, experiments of classification and recognition are reported to show how difficult is the final goal of this task.

3.1. Front-End Audio Features

This section is a summary of the front-end audio feature extraction method used in [18]. We extract 16 MFCCs (including C0) computed in 25ms frame size with a 10ms frame step and their Δ and $\Delta\Delta$. Due to the high variability of every acoustic concept, the fact that the segments are overlapped with speech and music, and the different devices used to record the video, a normalization of these features is needed. Trying to generalize the features, a cepstral mean normalization is computed over the whole video and the mean and standard deviation are computed over 1-second windows with an overlap of 0.75 seconds. Thus, the system uses 96 features (48 for the mean and 48 for the standard deviation of the $MFCC + \Delta + \Delta\Delta$ features) every 0.25 second.

3.2. Classification System

This experiment shows how difficult the task is. The goal of this experiment is the classification of a set of cut segments in one of the broad classes. The segments are overlapped with speech and music in the background in some cases. However, the classification is done with the five broad classes (without speech and music models) keeping the seven broad classes (with speech and music models) for the recognition task. The segments are extracted from the video database generating 13520 segments of different durations. Each concept is model as one state HMM/GMM with 256 Gaussians. Table 3 shows the results using the same subset of data to train and test. As it can be seen, the task is very difficult due to the high within-class variability of each concept. The system classified 71.1% of the segments correctly.

To test the system, a 4-fold cross-validation was performed using 3 folds to train the models and 1 fold to test. Table 4 shows the confusion matrix and how the classification rate is reduced compared with Table 3, classifying a 45.9% of the segments correctly. It can be seen that the Animal Noise and the Environmental Sounds are the concepts with a higher error rate

	CA	AN	RS	MN	ES
CA	**0.77**	0.04	0.04	0.06	0.09
AN	0.08	**0.80**	0.04	0.02	0.06
RS	0.08	0.04	**0.75**	0.05	0.08
MN	0.11	0.04	0.09	**0.61**	0.16
ES	0.12	0.09	0.08	0.07	**0.63**

Table 3: Confusion Matrix using the same set for train and test

	CA	AN	RS	MN	ES
CA	**0.61**	0.03	0.06	0.12	0.18
AN	0.18	**0.12**	0.20	0.19	0.31
RS	0.11	0.05	**0.45**	0.18	0.21
MN	0.17	0.02	0.16	**0.40**	0.25
ES	0.24	0.07	0.15	0.21	**0.33**

Table 4: Four Folds Cross-Validation Confusion Matrix

because both classes do not have enough data to train the models.

3.3. Recognition System

In the MED task, a recognition system is needed to be able to detect and classify the acoustic concepts related with the video. Due to the fact that most of the acoustic concepts are overlapped with speech and music, two extra models are needed to identify the segments in which there is not an acoustic concept. Also, these models can be useful to describe the video in the MED task. Using the same models trained for the classification task, a segmentation is executed over the whole video where the cut-segments were extracted for the classification system in previous subsection.

In this experiment, every concept (speech and music included) is modeled by a HMM/GMM of one state. The main difference is that a segmentation is produced when there are transitions between the models in the Viterbi algorithm. Table 5 shows the recognition result per concept independently of the segment duration. As it can be seen, Crowds and Repetitive Sounds have the better results in comparison with the Animal Noise or Environmental Sound because Crowd and Repetitive sounds were trained with more data than Animal Noise or Environmental Sound. The following sections show how the multimedia events related with the acoustic concepts Animal Noise or Environmental Sound have a poor detection rate because the models are not well-trained.

4. Acoustic concepts as features for MED

4.1. Methods

The purpose of the acoustic concepts recognition is to enable a video to be modeled by the acoustic concepts present in the video. For example, the ability to identify certain properties of the audio component that correlate strongly with crowd sounds and little with environmental sounds such as water might indicate the video takes place in a setting with large number of people present away from water and is therefore more likely to belong to certain video events (i.e. parade) than others (i.e. fishing). This section shows two different approaches using acoustic concepts to detect the multimedia event.

	CA	AN	RS	MN	ES	SP	MU
CA	**0.41**	0.03	0.03	0.04	0.04	0.20	0.23
AN	0.11	**0.01**	0.01	0.05	0.07	0.52	0.20
RS	0.07	0.02	**0.35**	0.09	0.09	0.16	0.20
MN	0.14	0.10	0.10	**0.26**	0.16	0.07	0.15
ES	0.23	0.02	0.07	0.05	**0.11**	0.12	0.13

Table 5: Segmentation Confusion Matrix

The first one is described in [15] and it is known as Segmental-GMM. Training the GMM for the seven selected acoustic concepts, a score vector is generated on fixed-length segments with each element in the vector corresponding to a posterior score for a GMM. As mentioned, we refer to these score vectors as Segmental-GMM feature vectors. In our experiments the segmental GMM vectors are 7-dimensional.

The second approach is known as Acoustic Concept Recognition (ACR) in which each concept is modeled as a HMM/GMM of one state. The main difference with the Segmental-GMM approach is that the segments are not fixed-length any more, and the segmentation is based on the transitions between the HMM models following the Viterbi algorithm. The score vector is the accumulated likelihood for each model. Therefore, a video is represented by a 7xK dimensional matrix with each column corresponding a different length segments.

In order to perform classification on the multimedia event level, we need to have features that are constant length independent of the video length. These constant-length features can then be used with the SVM classifier. The original video is currently represented by a 7xK matrix and is therefore not fixed-length. In this work, we represent a video with what we refer to as a *co-occurrence matrix* in which each element represents the probability that a pair of acoustic concepts occur in the video. This process is described in [15].

We performed a verification, also referred to as *one-against-all*, experiment for each of the five video event classes. For each video event, a given file is labeled as *in-class* or *out-of-class*. For example, for E004 we would perform the binary classification into *Wedding ceremony* and non-*Wedding ceremony*. We chose to perform classifications using support vector machines (SVMs) with a linear kernel. SVMs are commonly used for similar classification experiments due their simplicity and ability to model nonlinear decision boundaries using what is referred to as the 'kernel trick.'

4.2. Results

To measure the system performance results we use Detection Error Tradeoff (DET) curves, which are commonly used to show the tradeoff between the false alarm errors and missed detections. We generated the DET-curves in this paper with plotting software available from the NIST website [19]. From these curves, we also extracted the equal error rate (EER) as the the point where the probability of false alarm (pFA) is equal to the probability of a miss (pMiss). Since TRECVid MED 2011 simulates a retrieval task from wild videos in the internet, the assumption is that high miss rates can be tolerated in favor of low false alarm probabilities. Therefore, we use a benchmark to compares the number of misses at a given false alarm rate of 6%. The percentage of misses at a given false alarm rate is computed in a similar fashion to EER.

Figure 1 shows the DET curves for every acoustic event. The blue curves represents the performance of the Segmental-GMM approach, and the red curve represents the performance of the ACR approach. As it can be seen, the systems per-

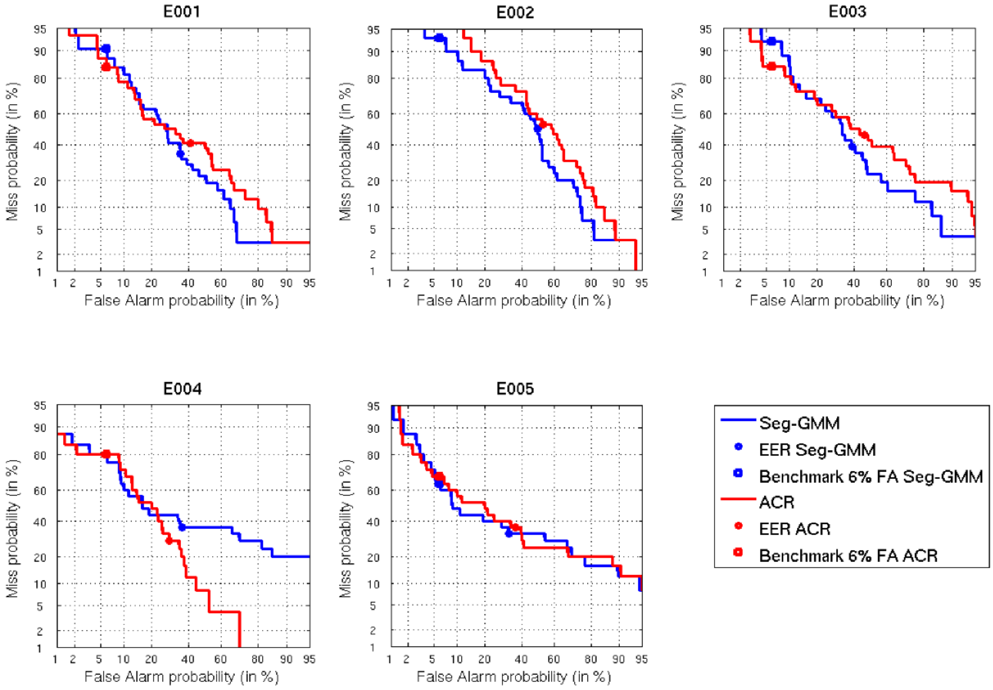

Figure 1: DET curves of Segmental-GMM approach versus ACR approach. The marks for EER and the benchmark for 6% of pFA are on the same curves

formance varies across video events. *Wedding ceremony* and *Woodworking project* show the best results while *Feeding an animal* and *Landing a fish* show the worst results. These behaviors are consistent with the previous results in section 3. It can be seen that the concepts *Animal sounds* and *Environmental sound* have the biggest error rate, and those concepts are more related with *Feeding an animal* and *Landing a fish* videos respectively. On the other hand, the concepts *Crowds and audience* and *Repetitive sounds* have the best results, and they are more related with *Wedding ceremony* and *Woodworking project* events respectively. Also, *Feeding an animal* and *Landing a fish* videos contain short bursts of sounds overlapping with a widely varying background noise, which make the detection much more difficult.

Table 6 shows the EER and the benchmark given a false alarm rate of 6% for both approaches. The EER is better using Segmental-GMM for almost all the events except for the event *Wedding ceremony*. However, the benchmark is better using ACR with the exception of E002 event where the model is poor trained and E005 where the difference between Segmental-GMM and ACR is not significant as can be seen in Figure 1.

	Segm-GMM		**ACR**	
	EER	BM-6%	EER	BM-6%
E001	**0.343**	0.906	0.406	**0.843**
E002	**0.500**	**0.933**	0.533	1.000
E003	**0.384**	0.923	0.461	**0.846**
E004	0.360	0.800	**0.280**	**0.800**
E005	**0.320**	**0.640**	0.360	0.680
Mean	**0.381**	0.840	0.408	**0.833**

Table 6: EER and Benchmark of 6% of pFA for segmental-GMM and ACR approaches

5. Conclusions

This paper shows a comparative study between different approaches to detect multimedia events using a set of videos provided in TRECVid 2011 evaluation. These approaches are based on the analysis of the audio of the videos, and they help to improve the detection accuracy of video analysis systems. The proposed approaches create features based on the likelihood of acoustic concepts that can happen in the multimedia event.

The first set of experiments shows the accuracy to classify and recognize the acoustic concepts. The videos of the TRECVid 2011 are downloaded from different sources in internet, so the audio of these videos has a lot of variability. The acoustic features that compensate the variability of the audio are the mean and the variance of MFCCs. However, training and testing over the same set of data provide a mean error rate of 30% as it was showed in Table 3. The concepts *Animal Sounds* and *Environmental Sound* have the highest error rate for all the systems and, therefore, the events related with these concepts (as *Feeding an animal* and *Landing a fish*) have the highest detection error rates for all the event detector approaches.

We create a baseline based on the approach proposed in [15]. This baseline is known as *Segmental-GMM* and it creates a feature vector with the likelihood of the acoustic concepts from a GMM model for every acoustic concept extracted every five seconds. The novelty proposed in this paper is to create an HMM-GMM model for every acoustic concept to be able to get a segmentation based on the transitions between the models. This solution is know as ACR and it shows a little improvement over the *Segmental-GMM* as a retrieval approach.

6. Acknowledgments

We would like to thank Stephanie Pancoast for her help with the annotations and baselines.

This work has been funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02 when the authors were at SRI International.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes non withstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7. References

[1] T. multimedia event detection 2011 evaluation, "http://www.nist.gov/itl/iad/mig/med11.cmf."

[2] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[3] C. Zieger, "An hmm based system for acoustic event detection," in *Multimodal Technologies for Perception of Humans*, 2008.

[4] Mostefa, Moreau, and Choukri, "The chil audiovisual corpus for lecture and meeting analysis inside smart rooms," in *Evaluation and Language Distribution Agency*, 2008.

[5] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia*, 2006.

[6] C. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, 2007.

[7] S. Tsakalidis, X. Zhuang, R. Hsiao, S. Wu, P. Natarajan, and R. Prasad, "Robust event detection from spoken content in consumer domain videos," in *Interspeech 2012*, 2012.

[8] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Interspeech 2012*, 2012.

[9] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech2012*, 2012.

[10] L. Li, "A novel violent videos classification scheme based on the bag of audio words features," in *International Journal of Computational Intelligence*, 2012.

[11] B. Byun, S. Kim, I.and Siniscalchi, and L. C.H., "Consumer-level multimedia event detection through unsupervised audio signal modeling," in *Interspeech 2012*, 2012.

[12] S. Chaudhuri, R. Singh, and R. Raj, "Exploiting temporal sequence structure for semantic analysis of multimedia," in *Interspeech 2012*, 2012.

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

[14] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and D. A., "Acoustic super models for large scale video event detection," in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*. ACM, 2011.

[15] S. Pancoast, M. Akbacak, and M. Sanchez, "Supervised acoustic concept extraction for multimedia event detection," in *ACM Multimedia Workshop*, 2012.

[16] Y. Jiang, X. Zeng, G. Ye, and S. Bhattacharya, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVID 2010*, 2010.

[17] J. Van Hout, M. Akbacak, D. Castan, E. Yeh, and M. Sanchez, "Extracting spoken and acoustic concepts for multimedia event detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013*, 2013.

[18] D. Castan, C. Vaquero, A. Ortega, and E. Lleida, "Hierarchical audio segmentation with hmm and factor analysis in broadcast news domain," in *Interspeech2011*, 2011.

[19] N. DETware V.2., "http://www.itl.nist.gov/iad/mig/tools/."