Indexing Multimedia Documents with Acoustic Concept Recognition Lattices

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 data is increasing every day and there is a growing demand for high-accuracy multimedia retrieval systems that go beyond retrieving simple events (e.g., detecting a sport video), to more specific and hard-to-detect events (e.g., a point in a tennis match). To retrieve these complex events, audio content features play an important role since they provide complementary information to image/video features. In this paper, we propose a novel approach where we employ an HMM-based acoustic concept recognition (ACR) system and convert resulting recognition lattices into acoustic concept indexes to represent multimedia audio content. Lattice indexes are created by extracting posterior-weighted N-gram counts from the ACR lattices and they are used as features in SVM-based classification for multimedia event detection (MED) task. We evaluate the proposed approach on the NIST 2011 TRECVID MED development set, which consists of user-generated videos from the internet. Proposed approach yields an Equal Error Rate (EER) of 31.6% on this acoustically challenging dataset (on a set of 5 video events) outperforming previously proposed supervised and unsupervised approaches on the same dataset (34.5% and 36.9% respectively).

Index Terms: Multimedia event detection (MED), acoustic concept recognition, lattice N-gram counts, acoustic concept indexes

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.

Audio concept extraction approaches explored under different multimedia retrieval and content analysis projects, such as *multimedia event detection* (MED), 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 ([1, 2]). Other unsupervised approaches are focused on segmenting the audio track, and clustering the segments to form atomic sound units and then word-like units [3, 4], or modeling the segments with i-vectors [5] or GMM super-vectors [6] which are methods borrowed from speaker identification. In the second group of approaches, audio concept/event models are trained using annotated data [7, 8]. For example, in [7], 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 [8], 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 [9] 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.

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

Here, we propose a novel approach to model multimedia audio content with a supervised acoustic concept extraction technique. First, we employ an HMM-based *acoustic concept recognition* (ACR) system to convert audio signal into a recognition lattice or what we refer to as *acoustic concept lattice*. Next, we create an acoustic concept index for each file from the ACR lattice by extracting posterior N-gram counts. This is inspired by previous work in language recognition systems where phonetic indexes are extracted for dialect identification task [10]. Finally, the acoustic concept indexes are used as features in SVM-based classification for multimedia event detection (MED) task. This approach is different from the previously mentioned supervised techniques [7, 8] in several ways. First, we do not use any fixed segmentation, but instead use recognition to extract acoustic concept segments dynamically. Another difference from [8] as opposed to using only broad class acoustic concepts (e.g., crowds sound or indoor/outdoor sounds), we also use more specific acoustic concepts (e.g., crowds cheering, crowds laughter). More importantly, in our approach soft-decisions for the acoustic concept extraction are used as MED

5 Broad Concepts	20 Specific Concepts	
	Air traffic	Individual applause
	Birds	Individual yells
Crowds/audience	Crowd applause	Large crowd
Animal sounds	Crowd cheers	Scraping-Sanding
Repetitive sounds	Crowd laughter	Sewing
Machine noise	Crowd yells	Skateboard
Environmental sound	Farm animals	Small party
	Ground traffic	Water running
	Hammer	Water splashing
	Wind	Home appliances

Table 2: Broad and Specific acoustic concepts

features via lattice-based representations to consider alternative recognition hypothesis creating rich representations to be used for MED task. Given the amount of variation in audio characteristics of user-submitted internet videos, this becomes critical since 1-best hard-decisions will most likely have high-error rates and this will degrade MED performance. And the last difference is that in our work context information is used (via N-gram representations) during MED modeling.

The paper is organized as follows: dataset and the acoustic concept annotations are described in Section 2. Section 3.1 shows the front-end features and the acoustic concept modeling used in this approach. Indexing acoustic concept recognition lattices and how they are used as features for MED task are explained in Section 3.2. In Section 4 a performance comparison between previously reported approaches and our proposed approach is presented as well. Finally, conclusions and future work are presented in Section 5.

2. Dataset and Acoustic Concept Annotations

2.1. TRECVID MED 2011 Dataset

The Text Retrieval Conferences Video Retrieval Evaluation (TRECVID) [11] 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 difference 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 [7] and [1] 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 Concept Annotations

Our goal is to describe the multimedia events using acoustic concept recognition output as high-level features. Therefore two sets of labels of acoustic concepts had been created in [7] where details on the acoustic concept annotation can be found. The first set has five labels representing five broad acoustic concepts. These broad concepts can be divided into more specific acoustic concepts trying to be more descriptive. Table 2 shows the two sets of acoustic concepts. We inherit the same annotation in our study with a little modification: these acoustic concepts have been extended with "*speech*" and "*music*" classes because most of the videos contain speech or music as the predominant audio. Furthermore, some acoustic concepts are overlapped with speech or music and sometimes they are barely audible in the background, and this presents a difficulty for extracting acoustic concept labels correctly.

3. System Overview

3.1. Front-End Features and Acoustic Concept Modeling

As we described in Section 2, a subset of data is used to train the acoustic concepts models. This set consists a total of 1536 videos (47 hours approximately) averaging 1.8 minutes per file.

As a front-end features we compute 16 Mel Frequency Cepstrum Coefficients (MFCCs) (including the zero-coefficient) in 25 ms frame size with a 10 ms frame step, their Δ and $\Delta\Delta$. Due to the high variability of the videos and the fact that the segments are overlapped with speech and music, 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 deviations are computed over 1 second windows with an overlap of 0.75 second. 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. Each concept is modeled as one state HMM/GMM with 256 Gaussians. During decoding, one-state HMM model per concept with a null grammar is used. We use the HTK HMM toolkit [12] to run acoustic concept recognition and produce lattices. Next subsection will present how we create acoustic concept indexes from recognition lattices.

3.2. Lattice-based Acoustic Concept Indexing

In our proposed approach, we extract the sequence of acoustic concepts and use their posterior-weighted counts as features to classify the target multimedia event. The main idea is that a sequence of acoustic concepts can be indicative of a specific multimedia event. This approach has been applied successfully to identify different languages [13] or different dialects [10] by using phonetic N-gram counts.

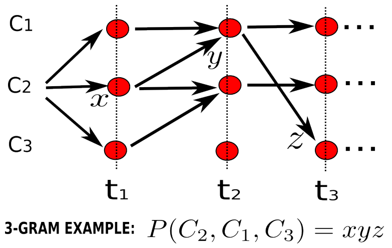

Figure 2: An example of 3-gram extraction from a sample acoustic concept recognition (ACR) lattice

Lattices represent alternative hypotheses resulting in richer representation compared to 1-best recognition output. These hypotheses can be seen as multiple paths with different likelihood for every node of the path. In our approach, each node represents an acoustic concept. The N-gram counts are the

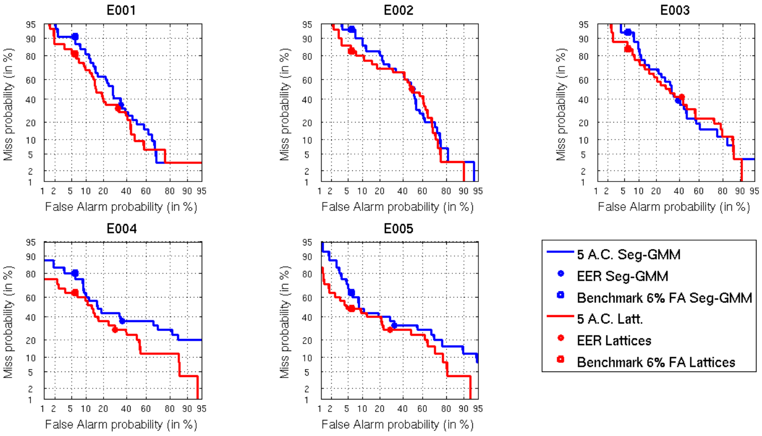

Figure 1: DET curves of Segmental-GMM and ACR-lattices approaches with 5 acoustic concepts. The marks for EER and the benchmark for 6% of pFA are on the same curves

accumulated likelihoods of co-ocurrence concepts as Figure 2 represents. A deeper explanation can be found in [14] where similar lattice-based indexing approach is presented to identify languages. We use the N-gram counts as a high-level feature vector to train a support vector machine (SVM) to detect the events. Therefore, we train five SVMs (one per event presented in Table 1) to determine if an event belongs to a class or not. Resulting feature vector is called *acoustic concept index* and SVM feature space, in this case, is the acoustic concept index of the event.

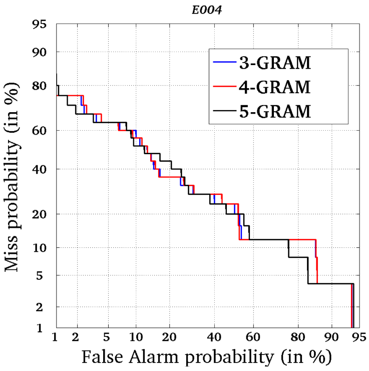

Figure 3: DET curves for 3-GRAM, 4-GRAM and 5-GRAM for the same event (E004). Increasing the N-gram size does not improve the MED results

4. Experimental Results

We evaluate the performance of lattice-based acoustic concept indexing in a multimedia event detection (MED) task. To measure the system performance, we use Detection Error Tradeoff (DET) curves, which are commonly used to show the tradeoff between the false alarm errors and missed detections. We generate the DET curves using [15]. We also report the equal error rate (EER). In a retrieval task, high miss rates can be tolerated in favor of low false alarm probabilities. Therefore, the benchmark (BM-6%) 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.

The acoustic concept HMM models described in Section 3.1 were used to generate recognition lattices across the train (*train-2*) and the test datasets. As an initial experiment, the SVM vectors were produced stacking 1-grams, 2-grams and 3-grams of the 5 broad acoustic concepts getting 520 vector dimensions. To get a sense of how well the lattice-based acoustic concept indexing approach (denoted as A.C. Latt.) performs, we compare the DET curves against segmental-GMM approach (denoted as A.C. seg-GMM in the plots) [7] for the same 5 acoustic concepts. This is shown in Figure 1. It can be seen how the DET curve for the ACR-lattices approach is below the segmental-GMM approach curve for most of the events. First part of the Table 3 shows the results are improved by employing ACR-lattices approach in terms of EER and a benchmark of 6% of pFA. The proposed approach, ACR-lattices, improve the detection in almost all the events. The table shows that ACR-lattices is the best performing approach for a retrieval system because it has the lowest pMiss for the benchmark of 6% of pFA. However, E002 and E003 still have the worst behavior due to the amount of data to train the acoustic concept models that appear in these video event categories.

The number of SVM vector dimensions grows exponentially as we increase the order of the N-grams. Figure 3 shows that increasing the SVM vector with the counts of 4-grams and 5-grams (getting 3656 and 25608 dimensions respectively) does not improve the results. The figure shows the curve for the event E004 because it is the best event in terms of EER. Although, the second part of the Table 3 shows the results for all the events and it can be seen that there is not improvement increasing the N-grams. It may be because the information entropy to detect the video events is located in the first orders of the N-grams.

Next, we compare the proposed ACR-lattices approach with other approaches outside this paper. We use the 20 acoustic concepts and the segmental-GMM approach used in [7]. We also compare proposed approach with an unsupervised approach which is a bag-of-audio-words (BoAW) [1]. For all of these approaches, same train and test sets are used. Because the segmental-GMM approach is evaluated with 20 acoustic concepts instead of the 5 broad acoustic concepts used in this paper, we extend the ACR-lattices approach with 20 acoustic concepts using 1 state HMM with 256 Gaussians for every acoustic concept as we did for the 5 broad acoustic concepts. The SVM vectors in this experiment use 3 grams because increasing order of N-gram further shows a little improvement in the detection performance, and also vector dimensions increase exponentially. For the BoAW approach, a codebook size of 1000 is used as

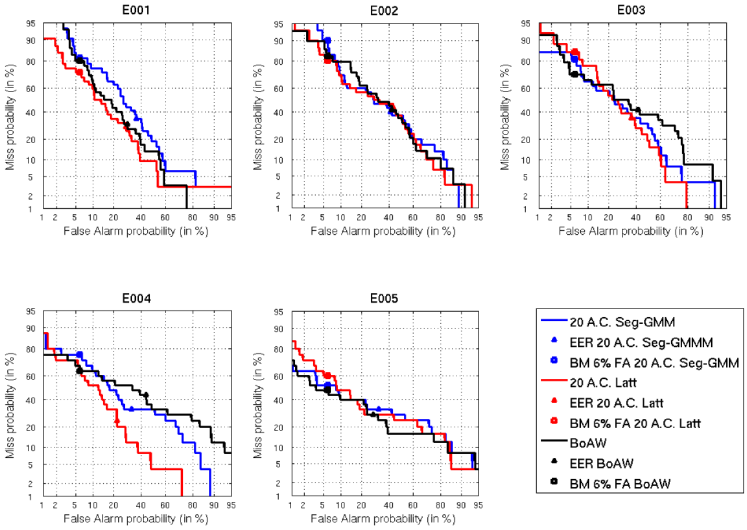

Figure 4: DET curves of segmental-GMM with 20 acoustic concepts, ACR-lattices with 20 acoustic concepts and BoAW approaches. The marks for EER and the benchmark for 6% of pFA are on the same curves.

System	EER/BM	E001	E002	E003	E004	E005	Avg.
5 A.C. SegGMM	EER	34	50	38	36	32	43.5
	BM-6%	90	93	84	80	64	84.0
5 A.C. Latt	EER	31	50	42	28	28	**35.9**
	BM-6%	81	83	84	64	48	**72.2**
System	EER/BM	E001	E002	E003	E004	E005	Avg.
3-GRAM Latt	EER	31	50	42	28	28	35.9
	BM-6%	81	83	84	64	48	72.2
4-GRAM. Latt	EER	31	46	42	28	28	35.2
	BM-6%	78	83	84	64	48	71.6
5-GRAM. Latt	EER	31	46	42	28	28	35.2
	BM-6%	78	83	84	64	48	71.6
Systems	EER/BM	E001	E002	E003	E004	E005	Avg.
20 A.C. SegGMM	EER	34	40	34	32	32	34.5
	BM-6%	81	90	80	76	52	75.9
20 A.C. Latt	EER	28	43	34	24	28	**31.6**
	BM-6%	71	80	84	64	60	72.0
BoAW	EER	30	41	41	44	28	36.9
	BM-6%	80	82	70	64	48	**69.1**

Table 3: EER and Benchmark of 6% of pFA for segmental-GMM (with 5 and 20 acoustic concepts), ACR-lattices (with 5 and 20 acoustic concepts and different N-grams), and BoAW approaches in percentage. Best performance numbers are highlighted.

this was found to yield the best BoAW results in [1].

Figure 4 shows the DET curves for these approaches. As can be seen, for almost all the events, the curve corresponding to ACR-lattices approach has better behavior than the BoAW and segmental-GMM curves. Also, the performance of the BoAW for events E002 and E003 is very bad even if the BoAW creates unsupervised clusters. That shows the difficulty of detecting that multimedia events using the audio of the video. Table 3 summarizes the EER and the pMiss for a benchmark of 6% of pFA. Also for these marks, the ACR-lattices approach shows very good behavior.

5. Conclusions and Future Work

We address the problem of modeling audio content for multimedia event detection task in user-submitted videos. We propose a novel approach where we employ an HMM-based acoustic concept recognition (ACR) system and convert resulting recognition lattices into acoustic concept indexes to represent multimedia audio content. Lattice indexes are created by extracting posterior-weighted N-gram counts from the ACR lattices and they are used as features in SVM-based classification for multimedia event detection (MED) task. We evaluate the proposed approach on the NIST 2011 TRECVID MED development set, which consists of user-generated videos from the internet. Proposed approach yields an EER of 31.6% on this acoustically challenging dataset on a set of 5 video events. We compared our proposed approach against a previously proposed supervised and unsupervised acoustic content modeling approaches such as bag-of-audio-words (BoAW) approach [1] and segmental-GMM approach[7]. In the future, we would like to use discriminative acoustic modeling techniques for acoustic concept recognition component to increase separation between confusable acoustic concepts. We also would like to explore other front-end features. Future work will also include combining ACR-lattices approach with other two techniques that are presented in this paper for better MED performance.

6. Acknowledgments

We thank Stephanie Pancoast for her help with her segmental-GMM and BoAW setups. This work has been partly funded by the national project TIN2011-28169-C05-02 and the European Union (FEDER) when the authors were working in 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] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech2012*, 2012.

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

[3] 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.

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] M. Akbacak, D. Vergyri, A. Stolcke, and S. N., "Effective arabic dialect classification using diverse phonotactic models," in *Interspeech*, 2011.

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

[12] H. Toolkit, "http://htk.eng.cam.ac.uk/."

[13] W. Campbell, F. Richardson, and D. Reynolds, "Language recognition with word lattices and support vector machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[14] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

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