Acoustic Super Models for Large Scale Video Event Detection

Robert Mertens
International Computer
Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
rmertens@icsi.berkeley.edu

Howard Lei
International Computer
Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
hlei@icsi.berkeley.edu

Luke Gottlieb
International Computer
Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
luke@icsi.berkeley.edu

Gerald Friedland
International Computer
Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
fractor@icsi.berkeley.edu

Ajay Divakaran
SRI International Sarnoff
201 Washington Road
Princeton, NJ 08540, USA
ajay.divakaran@sri.com

ABSTRACT

Given the exponential growth of videos published on the Internet, mechanisms for clustering, searching, and browsing large numbers of videos have become a major research area. More importantly, there is a demand for event detectors that go beyond the simple finding of objects but rather detect more abstract concepts, such as "feeding an animal" or a "wedding ceremony". This article presents an approach for event classification that enables searching for arbitrary events, including more abstract concepts, in found video collections based on the analysis of the audio track. The approach does not rely on speech processing, and is language-indepent, instead it generates models for a set of example query videos using a mixture of two types of audio features: Linear-Frequency Cepstral Coefficients and Modulation Spectrogram Features. This approach can be used in complement with video analysis and requires no domain specific tagging. Application of the approach to the TRECVid MED 2011 development set, which consists of more than 4000 random "wild" videos from the Internet, has shown a detection accuracy of 64 % including those videos which do not contain an audio track.

Categories and Subject Descriptors

H5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*; H3.1 [**Information Storage and Retrieval**]: Indexing Methods

J-MRE'11, November 30, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0996-7/11/11 ...$10.00.

General Terms

Algorithms, Performance, Experimentation

Keywords

TRECVid, Multimedia Event Detection, Audio Processing

1. INTRODUCTION

Data-driven event classification can be considered more flexible than symbolic or semi-symbolic approaches that require manual crafting of event descriptions. In most cases, symbolic event descriptions also require manual tagging of vast amounts of training data (e.g., in platforms like LabelMe [5]) in order to learn how to detect the elements found in a symbolic event description.

Consider for instance a description of a wedding, one event class found in the TRECVid MED 2011 data set. A simple description would probably include a couple (two people – in most cases one male "groom" and one female "bride") being wed by a third person (priest or official) in the presence of an audience (group of people watching). In order to enable a system to search for the elements of this description in a video, it must first be able to detect elements fitting the concept of bride, groom, priest, etc. This in turn requires these concepts to be somehow modeled in the system, usually requiring a vast amount of manual tagging. The situation is aggravated by the fact that tagging is mostly domain specific. Hence a new kind of event description often requires a significant amount of manual tagging in order to be usable for a search query. This problem can also arise within descriptions of the same event as a single concept might be represented by two or three different topics (e.g. Surfing, skateboarding and skiing all are classified as board tricks).

Another drawback of using manual event descriptions is that they are designed by humans but are to be used by computer algorithms. The point in making this distinction is that computer algorithms might work in a way that fundamentally differs from the way that human perception works. Hence two video or, for that matter, two audio files mthat look alike to a human might be rated as completely different by an algorithm. The same holds true in the opposite case, as two video or audio files rated by an algorithm as similar might show different concepts to a human. However, as the actual search is conducted by a computer, the feature values that discriminate events best in a computer algorithm should be focused on. An important advantage of data-driven approaches in this context is that multiple examples for one event class are fed into a system and that the algorithm filters out those features that best discriminate the instances with respect to the given classes, thus in a way translating the human concept into a specification that can be interpreted by a machine. One should keep in mind, however, that this translation just coincides with concepts as represented by the training data. So selection of the training data is crucial. Selecting the right training data still requires less effort than does hand coding event descriptions since hand coding involves coding at multiple levels starting at the semantic level down to connecting semantic concepts with feature level concepts. While these arguments speak in favor of data-driven approaches for event detection, their applicability still has to be proven. Moreover, current computational constraints make it difficult to extract and compare visual features for 25 frames or more per second in a selection of thousands of high definition videos.

This article describes the first approach of acoustic data-driven event detection for large scale retrieval tested on the TRECVid MED 2011 development data set. While approaches for earlier TRECVid challenges have in some instances used speech recognition or detection of preselected audio sub-concepts (see Section 2), audio features have never been used for TRECVid in a purely data-driven way. The results presented in this paper show that learning based on arbitrary low level audio features can be used to detect high level concepts like those given in TRECVid MED 2011. This insight is not only valuable in itself, it also suggests a role for low level audio features as an option in multimodal concept detection thus augmenting present concept detection systems. The paper is organized as follows. Section 2 gives an overview of related work in both the field of acoustic and visual multimedia analysis as well as multimedia event detection. Section 3 discusses the TRECVid MED 2011 data set before Section 4 describes the algorithms employed. An evaluation of the approach is presented in Section 5 before Section 6 concludes the paper with a discussion of evaluation results as well as future work.

2. RELATED WORK

There is a wealth of related work in multimedia content analysis, especially video analysis. A comprehensive description of the related work would easily exceed the page limits of this paper. Therefore, we concentrate on surveying only parts of the most relevant work. Very few papers take into account audio, and most importantly most work concentrates on finding small events or objects rather than entire concepts. Very good summaries are provided by a 2006 article by Lew, Sebe, Djeraba, and Jain [6] and a 2009 article by Snoek and Worring [12]. In [9], for example, the authors present a content-adaptive audio texture based method to segment video into audio scenes. A similar idea using a different approach was proposed in [1], where an unsupervised method for repeated sequence detection in TV broadcast streams was presented. The approaches would be very interesting to try on the data we used. However, the work mainly concentrates on scene transition markers and thus works mostly for professionally produced content. In [6] the possibilities of using multimodal analysis including audio are emphasized as a major research challenge.

The TRECVid evaluation [6], organized on a year-by-year basis by the US National Institute of Standards and Technologies (NIST), investigates mostly visual event detection on broadcast videos [8]. The task is to detect concepts like “a person applauding” or “a person riding a bicycle”. However, while approaches for earlier TRECVid challenges have in some instances used speech recognition, audio has mostly been ignored, with the exception of the detection of preselected audio sub-concepts such as “outdoor urban”, “outdoor rural” or “indoor quiet” described in [4]. Many visual methods developed in the community are related to the research presented here and could be used to complement it. The Informedia project’s [15] basic goal is to achieve machine understanding of video and film media, including all aspects of search, retrieval, visualization and secularization in both contemporaneous and archival content collections. However, again, audio is not explored beyond speech recognition. TRECVid’s counterpart, the NIST Rich Transcription (RT) [7] evaluation, focuses on acoustic methods for transcribing multimedia content. The evaluation is currently focusing on meeting data, but previous evaluations included broadcast news from radio and television.

3. DATA SET

The TRECVid 2011 MED dataset is different from the original TRECVid dataset. The MED dataset is comprised of “found videos”, i.e. consumer-produced videos downloaded from various social networking sites. Most videos are very short (a couple of minutes) and not produced professionally. The query sets (so-called event kits) are comprised of fifteen categories, which we used for training our system, with only five of those categories available for testing purposes during our study. The event kits consist of a total of 2040 videos and the test set of a total of 4251 videos. The five event categories which are available in the test set are “attempting a board trick”, “feeding an animal”, “landing a fish”, “wedding ceremony”, and “working on a woodworking project”; the remainder of the videos in the test set are random videos not belonging to any of the event categories.

The number of videos in each category for train and test is available in Table 1. The contents of these videos is highly variant, for example, the concept “attempting a board trick” includes people skateboarding, snowboarding and surfing, while the “wedding ceremony” varies from a traditional catholic mass, to a Hindi ceremony, to home-made music videos.

4. SYSTEM OVERVIEW

Our system is derived from a GMM-UBM speaker recognition system [10] using C0–C19 Linear Frequency Cepstral Coefficients (LFCC) features with 25 ms windows and 10 ms stepsize, along with deltas and double-deltas (60 dimensions total) as well as Modulation Spectrogram (MSG) features.

Mel Frequency Cepstral Coefficents (MFCCs), which are a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency, are commonly used as features in speech recognition systems. LFCCs use the same transformation, except for the Mel scale. For event

Table 1: Number and Types of Videos in Training and Testset.

Category	Description	Train Data	Test Data
E001	Board Trick	160	111
E002	Feeding Animal	160	111
E003	Landing Fish	122	86
E004	Wedding	128	88
E005	Woodworking	142	100
E006	Birthday Party	173	0
E007	Changing Tire	110	0
E008	Flash Mob	173	0
E009	Vehicle Unstuck	131	0
E010	Grooming animal	136	0
E011	Make a Sandwich	111	0
E012	Parade	134	0
E013	Parkour	108	0
E014	Repair Appliance	123	0
E015	Sewing	116	0
Other	Random other	N/A	3755

classification we have found that using LFCCs gives simlar results and therefore Mel scaling is not needed. The Modulation Spectrogram provides an alternative and complementary representation of the speech signal with a focus on temporal structure [14] as it represents a filtered version of the spectrogram of a sound signal. The spectrogram of the signal is computed using an FFT with a step size of 10 ms and an analysis window of 25 ms. In contrast to LFCC features, where for each frame the DCT coefficients of the log-FFT amplitudes are computed, MSG analyzes the spectrogram using 18 bands from 0 to 8 KHz, filtering the resulting 18 temporal signals with two different filters in the Hz range: a 0-8 Hz filter and an 8-16 Hz filter. For each frame, the MSG features capture the low-pass and band-pass behavior of the spectrogram of the signal within each of the 18 subbands, resulting in a total of 36 features per frame. In contrast to LFCCs, the modulation spectrogram provides information about longer temporal phenomena as it uses 0.21 seconds of analysis to extract the features. Thus, we expect that, jointly with LFCCs, this representation of the spectrum of the signal will be richer and perform better. We normalize all features per file by the mean and standard deviation of that file. Audio files with features for which this normalization method produces poor feature values (i.e. having a low standard deviation may produce extraordinarily high feature values) are ignored. The GMM-UBM system is a widely-used approach to speaker recognition, and is easy to implement. Specifically, for each audio track, a set of features is extracted and one Gaussian Mixture Model (GMM) is trained for each acoustic event, using features from all its videos. In other words, several hours of videos are trained into one GMM, disregarding the underlying sub-event structure of the audio. Since the models therefore represent an entire event kit and not only an atomic event, we call them "super-models". To increase the salience of event specific feature values, we apply a MAP (maximum a posteriori) adaptation from a universal background GMM model (UBM), which is trained using LFCC/MSG features from all audio of all acoustic events in the training set (i.e. event-independent audio) [10]. For testing, the log-likelihood of LFCC and MSG features from each test audio are computed using the pre-trained GMM models of each acoustic event. Figure 1 illustrates the GMM-UBM system.

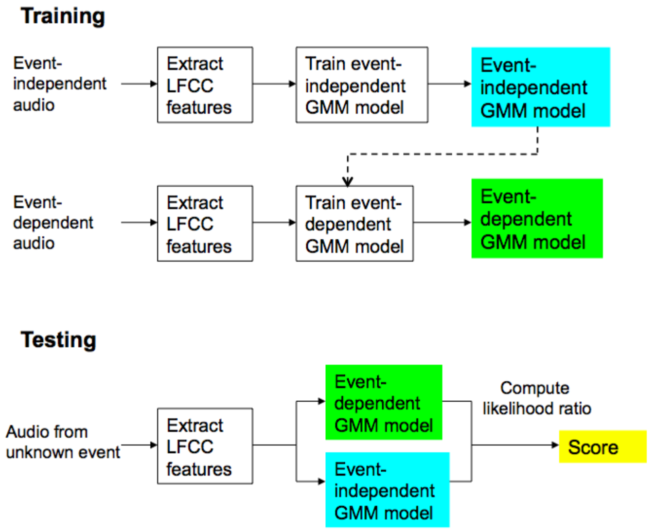
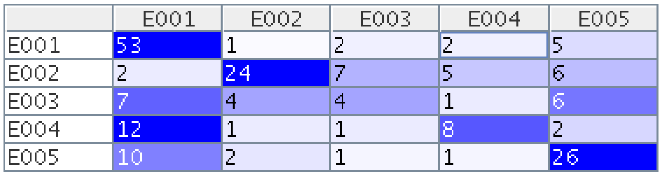


Figure 1: Simplified overview of GMM-UBM acoustic event detection system.

In order to combine LFCC features and MSG features on the model level, separate GMM-UBM systems are trained for LFCC and MSG. Scores from both systems are then combined to compute the overall score. The scores are weighted with the heuristically determined factors 6 for LFCC and 31 for MSG. A total of 128 mixtures are used for each GMM in the MSG system and a total of 1024 mixtures are used in the LFCC system. A likelihood score for each test audio as a match to each of the acoustic events is obtained as follows: Scores for which the acoustic event of the test video matches the event of the GMM model are known as true trial scores; scores for which the acoustic events do not match are known as false trial scores. During scoring, a threshold is established for distinguishing the true trial scores from the false trial scores. The system performance is based on EER, which is the false alarm rate (percentage of false trial scores above the threshold) and miss rate (percentage of true trial scores below the threshold) at a threshold where the two rates are equal. 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 the benchmark 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. A measure of accuracy can also be obtained by assigning to each test audio the event model with the highest score, and determining whether the assigned event model represents the actual event of the test audio.

The open-source ALIZE speaker recognition system implementation is used [2], and the 60-dimensional LFCC features as well as the 36-dimensional MSG features are obtained via HTK [3]. The experimental setup is designed to account for the influence of priors and data sparsity, as practiced in the US National Institute of Standards and Technology's (NIST) speaker recognition evaluations for many years [13].



	E001	E002	E003	E004	E005
E001	53	1	2	2	5
E002	2	24	7	5	6
E003	7	4	4	1	6
E004	12	1	1	8	2
E005	10	2	1	1	26

Figure 2: Confusion Matrix for E001–E005 as described in Section 5.

5. EXPERIMENTAL RESULTS

The experiments were conducted using the 4251 videos of the TRECVid MED 2011 open development set described in the previous section and using the system described in Section 4. A total 94 videos do not contain a soundtrack which are classified as "other". Since all of these files were actually part of the "other" category, they did not have any impact on the result. At a fixed false-alarm rate of 6 %, an unweighted model combination (weighting 1:1) results in a miss probability of 80.8 %. With a heuristically determined weighting of 6:31, a miss probability of 79.4 % was achieved. This is equivalent to an Equal Error Rate of 36 %. In the computation of the miss rate, a number of 8 files from the target set for which no features could be generated in a way they were usable for the system is also accounted for. These video files contained a soundtrack but had no audible sound.

It should be noted that this value was achieved with audio analysis only, so combination with visual features is expected to increase the overall performance. The audio based concept detection worked best on E001 and E005, the "board trick" and the "woodworking" scenario. Above the threshold score, the system correctly identified 30 videos for E001, 6 for E002, 7 for E004 and 24 for E005. Figure 5 shows the confusion matrix without the 6 % threshold thus showing the winning classifier for all events, even if the winning classifier's score was below the threshold. The DET curve for the evaluation run is depicted in Figure 5. The DET curve shows the relation of miss probability and false alarm probability for the experimental run. Both training and testing on over 6000 videos could be finished in about 10 hours on an 8-core commodity PC.

5.1 Analysis

Our system achieves the most promising results in the "attempting a board trick" and "working on a woodworking project" categories. Therefore, we decided to perform a more in-depth human analysis of these videos, with the hope of understanding what allowed the system to distinguish them. The five most common acoustic events in the "attempting a board trick" category according to human annotators were: Post-production music, speech, thumps (of a board landing on a surface), skateboard wheels on concrete, and slow motion distortion. A human analysis of the successfully categorized and miscategorized "board trick" videos revealed that the great majority of successfully categorized videos contained high tempo post-production music, and the majority of miscategorized videos had virtually no music at all. The sounds most commonly associated with board tricks (by humans) seem to have little bearing on successful categorization by our system. Similarly, in the "Woodworking" category our top-five acoustic events were as follows: Speech, Knife Scraping, Power Lathe, Hammer, and Table Saw. Speech was by far the most common sound in these videos, occurring in approximately 2/3 of the videos which we studied. Woodworking videos often have an individual clearly speaking to the camera, explaining technique. Our analysis of successfully classified videos suggests that the system is detecting speech of this type, as most of them have this layout. The actual sounds of woodworking seem to have little weight in our current system's classification of this type of video. An analysis of the top-ten videos classified as woodworking from the "other" category showed rather unsurprising results based upon our previous analysis: they were mostly "instructional" videos in which speakers used an instructional voice, consisting of a single person speaking directly to the camera. A few amusing outliers occurred, notably HVC328564, which was of people breaking wooden boards in a karate class, which one could argue is actually "woodworking" (albeit unusual). Overall, however, we interpret this as further indication that machines find other features discriminative than humans, which is again supporting the hypothesis for a data-driven approach.

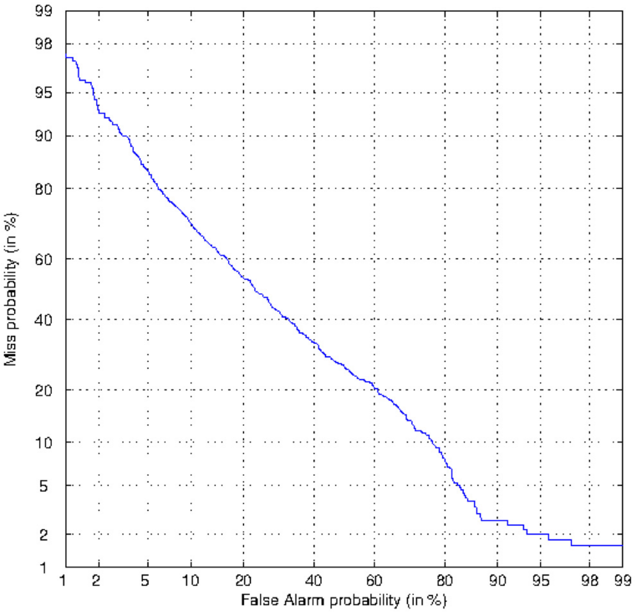


Figure 3: DET curve of results as described in Section 5.

5.2 Limits of the Approach

One shortcoming of the data-driven approach is that it is hard to explain why a video was classified as belonging to a certain event class and why another was not. It might for instance also be possible that the system is trained on too much noise i.e., it is overfitted to the current training set. Even though this is mitigated by training a background model, there is no way to analyze which sub-classes of sounds are learned by the system. While this fact is to a degree based on the nature of data-driven concept detection, the authors plan to address it in future extensions, for example by exploring the possibility of combining sub-event detection with the proposed approach. Another shortcoming that is shared with most other approaches is parameter sensitivity.

For example, initially, LFCC features from 280 files were not usable by the system. A later analysis has shown that LFCC features were generated using a 300 Hz low pass filter and a 3000 Hz high pass filter, a standard setting for many acoustic recognition tasks. After changing the high pass filter to 6000 Hz the number of unusable feature files dropped to 7 files. First experiments with LFCC in the new frequency range (tested with 128 Gaussians) have also indicated that using this broader range of frequency improves detection results. A simple 1:1 combination of the broader frequency LFCC and MSG has resulted in a miss probability of 78.4 % at a false alarm rate of 6 %. Note that the discussion of positively identified files in the previous section does not take this extended frequency range into account. We assume that the modification has also changed the type of sounds that are detected.

6. CONCLUSION

The two obvious media that can assist in video event detection are the visual and acoustic modalities. Traditional vision-based methods, however, mostly rely on analysis techniques that involve object recognition. This fact makes visual approaches highly domain dependent as the trained classifiers need to be trained based on human annotation. This is impractical with ever-changing sample queries over found consumer-produced data. Data-driven approaches for image analysis, such as "bag of words" approaches are highly computationally inefficient as they require the analysis of two to three orders of magnitude more data than audio-based analysis techniques[1]. Therefore, in order to explore the applicability of data-driven approaches to event detection, the approach presented in this paper is based on audio features only. The features used are LFCC and MSG. They are combined on a model level to combine the classification strengths of both features. The evaluation discussed in this paper has shown that a purely data driven approach for video concept detection can yield results that are comparable to visual approaches. Application of the approach to the TRECVid MED 2011 development set, which consists of more than 4000 random "wild" videos from the Internet, has shown a detection accuracy of 64 % which includes videos that do not contain an audio track. This is even more surprising as these results have been achieved with a purely acoustic approach and the challenge calls for multimodal approaches. The system presented in this paper is designed to be part of a multimodal event detection framework that will incorporate visual concept detection, OCR, audio transcripts and semantic analysis. It is expected that the combination of these different kinds of classifiers will lead to even better results.

Future work will include experiments with other feature sets and different frequency ranges. We also plan to address the fact that different classifiers perform differently on an event-class by event-class basis. In order to tackle the shortcoming that results can not be explained intuitively to the user, we explore the use of higher level audio features such as described in [11].

7. ACKNOWLEDGMENTS

This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding 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.

8. REFERENCES

[1] S. Berrani, G. Manson, and P. Lechat. A non-supervised approach for repeated sequence detection in TV broadcast streams. *Signal Processing: Image Communication*, 23(7):525–537, 2008.

[2] J.-F. Bonastre, F. Wils, and S. Meignier. Alize, a free toolkit for speaker recognition. In *ICASSP'05, IEEE*, Philadelphia, PA (USA), March, 22 2005.

[3] Hmm toolkit (htk). `http://htk.eng.cam.ac.uk`.

[4] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[5] LABELME. `http://labelme.csail.mit.edu/`.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[7] NIST Rich Transcription evaluation. `http://www.itl.nist.gov/iad/mig//tests/rt`.

[8] NIST TRECVid evaluation. `http://www-nlpir.nist.gov/projects/trecvid/`.

[9] F. Niu, N. Goela, A. Divakaran, and M. Abdel-Mottaleb. Audio scene segmentation for video with generic content. In *Proceedings of SPIE*, volume 6820, page 68200S, 2008.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000.

[11] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. On the use of audio events for improving video scene segmentation. In *WIAMIS 2010*, pages 1 –4, 2010.

[12] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundamental Trends in Information Retrieval*, 2(4):215–322, 2009.

[13] Nist speaker recognition evaluation. `http://www.itl.nist.gov/iad/mig/tests/sre`.

[14] O. Vinyals and G. Friedland. Modulation spectrogram features for speaker diarization. In *Proceedings of the 9th International Conference of the ISCA*, pages 630–633. Interspeech 2008, 2008.

[15] H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52, 1996.

[1] Remember that content analysis is usually performed by extracting features from raw, i.e. uncompressed, video. A second of mono audio sampled at 44.1 kHz with 16 bit resolution is 88 kB while a second of low-resolution video, say (320 × 240 RGB-pixels at 25 frames per second) is 5.76 MB.