INTERSPEECH 2009 BRIGHTON

Detecting Audio Events for Semantic Video Search

M. Bugalho[1 2], *J. Portêlo*[1], *I. Trancoso*[1 2], *T. Pellegrini*[1], *A. Abad*[1]

[1]INESC-ID Lisboa
[2]Instituto Superior Técnico, Lisboa, Portugal
Isabel.Trancoso@inesc-id.pt

Abstract

This paper describes our work on audio event detection, one of our tasks in the European project VIDIVIDEO. Preliminary experiments with a small corpus of sound effects have shown the potential of this type of corpus for training purposes. This paper describes our experiments with SVM classifiers, and different features, using a 290-hour corpus of sound effects, which allowed us to build detectors for almost 50 semantic concepts. Although the performance of these detectors on the development set is quite good (achieving an average F-measure of 0.87), preliminary experiments on documentaries and films showed that the task is much harder in real-life videos, which so often include overlapping audio events.

Index Terms: event detection, audio segmentation

1. Introduction

The framework for this work is the European project VIDIVIDEO, whose goal is to boost the performance of video search engines by forming a 1000 element thesaurus. When searching for semantic concepts, many may have associated audio cues. The purpose of this work is to explore these cues by audio event detection (AED). For the sake of space, the current paper does not cover the cues typically associated with audio segmentation: speech/non-speech detection, gender detection, speaker clustering, etc. For a review of our work on this topic, see [1].

Some audio events are associated with human voice, such as crying, laughing, coughing, and sighing. This paper does not cover this category of events either. In fact, we only focus on audio events not associated with human voice, and assume that the search for them is only made in segments which the speech/non-speech module has previously classified as non-speech.

Audio Events Detection is a relatively new research area with ambitious goals. Typical AED frameworks are composed of at least two parts: feature extraction and audio event inference. The feature extraction process deals with different type of features, such as: total spectral power, sub-band power, brightness, bandwidth, spectral envelope, flatness, centroid, and spread, ZCR (Zero Crossing Rate), MFCC (Mel-Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction), pitch frequency, harmonic spectral centroid, deviation, spread and variation, etc. Brightness and bandwidth are, respectively, the first and second order statistics of the spectrogram, and they roughly measure the timbre quality of the sound. They are very similar to the MPEG features centroid and spread, respectively, but use a linear frequency, instead of a logarithmic one.

Some of these features are common to the audio segmentation and speech recognition modules, others to music information retrieval (MIR), etc. Due to the large amount of features that can be extracted, considering them all can lead to lengthy training processes due to slow convergence of the classification algorithms. In this situation, it is common practice to use feature reduction techniques like Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), which map the features into a new vector space where the greatest variance by any projection of the data lies on the first coordinate, the second greatest variance lies on the second coordinate, and so on.

In the inference process, various machine learning methods are used to provide a final classification of the audio events such as rule-based approaches (RB) [2], Gaussian mixture models (GMMs) [3] [4] [5] , Support Vector Machines (SVMs) [4] [6] [5], and Bayesian Networks [7]. Although in our previous work we attempted to explore the time structure of audio events by using HMMs [8], the current paper covers only recent experiments with SVMs. In spite of the fact that multiple-class classifiers may potentially yield better results than one-against-all classifiers, we chose to build the latter, as this approach allows an easy extension to new semantic concepts.

Given the unavailability of a corpus labeled in terms of audio events, we used a sound effect corpus for training. The potential of this type of corpus was proved in early experiments with a small pilot corpus [8]. The extended training corpus and the small evaluation corpus of documentaries and movies will be described in section 2, together with the evaluation metrics. The next sections describe our multiple experiments with one-against-all SVM classifiers, feature analysis and dimensionality reduction. Finally, section 7 presents the main conclusions and future plans.

2. Corpora

Manual labeling of audio events is an extremely morose task, specially if one attempts to mark a relative large number of audio events (e.g. one hundred), in video clips where such events occur simultaneously with speech, with music, with each other, and with relative amplitudes which may make them very audible or barely audible. The need to avoid (or at least reduce) this morose task was the main motivation for adopting as the AED training set, a large corpus of sound effects, provided by B&G, one of the partners of the project, as each file typically contains a single type of sound.

The current sound effect corpus has approximately 18,700 files with an estimated total duration of 289.6h. It includes enough training material for over 100 different audio semantic concepts, but so far we have only considered the 47 most represented concepts in the corpus. The selection of the training/development files for each concept was based on the audio quality of the files (no background noise, no mixture of audio events, in order to avoid the need for manual labels), and involved the analysis of 7,800 files.

The list of concepts is presented in Table 1, together with

the number of files and corresponding duration that were used as training/development corpus for each classifier. Most of the files have a sampling rate of 44.1kHz, although many were recorded with a lower bandwidth (<10kHz).

In order to test the one-against-all detectors in a *real life* situation, we manually labeled a number of movies, documentaries (DOC), talk shows (TS) and broadcast news (BN) that were likely to contain this list of audio events. This *real life* evaluation corpus covers a very limited number of audio events.

2.1. Evaluation metrics

The development experiments described in this paper will be assessed in terms of the well-known F-measure. However, the experiments with the evaluation set will be assessed both in terms of the ratio (pr_p) of true positives (tp) over total number of positives (p), and the ratio (pr_n) of true negatives (tn) over total number of negatives (n). In this way, one can take into account the very low number of positive examples for each concept in the whole videos. The detection performance is frame-based, but classification results in the test set are smoothed over time.

Results for the evaluation set will also be presented in terms of the average precision ranking measure adopted by the consortium. This measure not only allows us to evaluate the degree of importance attributed to each of the retrieved audio events, but also combines the precision and recall into a single value. The measure also has the added value of being event-based and not frame-based. The value is proportional to the area under the precision and recall curve. Average precision is defined as:

$$AP = \frac{1}{|R|} \sum_{k=1}^{|R|} l_k \frac{|R \bigcap M_k|}{k} \quad (1)$$

where R is the complete set of audio events in the test set, M_k is the ranked list of the top k predicted events and l_k is 1 if the k was correctly predicted and 0 otherwise. An event is considered to be correctly predicted if more then 50% of the predicted label is contained in the true label or vice versa. Large events are divided in events of 5 second duration to correctly account for event size in the score.

Average precision is largely used for search tasks while F-measure is more suitable for large automated processes where a higher result precision is required. Both metrics are presented in the results.

3. Baseline SVM classifiers

The sound effect corpus was used to train one-against-all detectors for each concept. Our initial set of detectors was SVM-based, and the experiments were made using the LIBSVM toolkit [10]. Preliminary experiments [8] [9] compared the performance of a limited set of features: PLP or MFCC coefficients (19 + energy + deltas), ZCR, brightness, and bandwidth.

Around 90 files were used as negative examples for each concept, of which an average of 31 were used as the development set. As a starting point, analysis windows of 0.5s with 0.25s overlap were adopted. Three different kernels were considered for the SVM (linear, polynomial and radial basis function (RBF)). The F-measure results were generally very good (above 0.8) with the RBF kernel for the 47 concepts, as shown in the last column of Table 1. For six concepts, however, the polynomial kernel performed slightly better. The worst results (below 0.7) were obtained with Door, Fireworks, Hammering, and Saw_Manual. The difference between the performance of MFCC and PLP coefficients was not significant.

Concept	#Files	Duration	F-m
airplane jet	26	1210.2	0.82
airplane propeller	58	2523.3	0.81
applause	30	1308.4	0.99
bell electric	34	704.2	0.72
bell mechanic	117	4669.5	0.85
big cat	91	3038.5	0.89
bird	93	6339.8	0.91
bus	34	2736.2	0.90
buzzer	23	503.3	0.72
car	97	3722.2	0.95
cat	40	1110.6	0.81
chicken	16	434.6	0.86
cow	24	692.8	0.71
digital beep	38	970.1	0.88
dog	45	1860.5	0.95
door	113	1059.8	0.50
explosion	43	672.1	0.89
fire	53	4706.6	0.94
firework	22	692.9	0.46
frog toad	50	2775.0	0.92
glass	58	977.3	0.87
gunshot heavy	37	972.4	0.93
gunshot light	110	2435.4	0.87
hammer	45	1469.9	0.67
helicopter	26	1298.5	0.82
horn	80	1089.5	0.94
horse	85	3311.0	0.95
insect buzz	28	1823.5	0.98
insect chirp	33	3267.2	0.99
motorcycle	132	7784.0	0.94
pig	33	1490.0	0.74
rattlesnake	32	773.6	0.99
saw electric	38	1290.7	0.88
saw manual	24	887.8	0.63
sheep	33	1602.8	0.91
siren	47	1133.1	0.90
telephone analogic	17	562.3	0.97
telephone digital	14	337.5	0.91
thunder	52	1941.2	0.98
traffic	32	4396.9	0.91
train	82	4895.5	0.89
typing	32	2434.2	0.95
walking hard	93	4014.3	0.93
walking soft	86	4421.7	0.93
water	72	6147.7	0.98
whistle	46	601.3	0.84
wolf howling	31	1006.8	0.94

Table 1: Number and total duration of files for each of the 47 concepts, together with F-measure for the development set.

3.1. Results on the evaluation corpus

The results obtained on the evaluation corpus using the best combination of features and RBF kernel parameters on the development set are shown in Table 2. These results confirm that detecting audio events in real life data is much more challenging than the classification of isolated events. The worse performance may be due to the fact that audio events almost never occur separately, being corrupted by music, speech, background noise and/or other audio events.

Concept	Test file	pr_p	pr_n	F-m	AP
applause	TS1	0.29	0.98	0.44	0.33
	TS2	0.26	0.99	0.42	0.33
bird	DOC1	1.00	0.82	0.05	1.00
	DOC2	0.04	0.72	0.01	0.00
	DOC3	1.00	0.74	0.18	1.00
dog	DOC4	0.62	0.95	0.65	0.85
	DOC5	0.96	0.73	0.65	0.40
siren	007-AViewToAKill	0.33	0.96	0.29	0.12
	DieHard4	0.49	0.94	0.05	0.02
	BN1	0.21	0.97	0.18	0.02
telephone	TheMatrix	0.68	0.99	0.73	0.58
analogic	TheAviator	0.76	0.99	0.67	0.67
water	DOC6	0.45	0.94	0.15	0.08

Table 2: SVM results for the evaluation set (mean AP=0.415).

4. Feature analysis

A feature analysis was performed using two techniques: Information Gain and Correlation-based Feature Subset Selection [11] (CFS). The Weka software [12] was used to perform this analysis whose goal was to determine which features are important to distinguish an audio event from the "world".

The information gain, given by equation 2 (where H is the entropy), is a commonly used technique to measure the worth of a feature (Ft) to discriminate between classes (Cl). However, information gain does not consider the fact that two features might provide the same information.

$$InfoGain(Cl, Ft) = H(Cl) - H(Cl|Ft) \qquad (2)$$

To evaluate the overall worth of a set of features we have used the CFS algorithm. The CFS evaluates a set of features by considering the individual predictive ability of each feature (correlation between feature and class) along with the degree of redundancy between them (correlation between features). The heuristic value is given by equation 3, where $\overline{C_{cf}}$ and $\overline{C_{ff}}$ are the mean feature/class correlation and mean feature intercorrelation, respectively.

$$CFS_s = \frac{k \times \overline{C_{cf}}}{\sqrt{k + k(k-1) \times \overline{C_{ff}}}} \qquad (3)$$

The set of 79 features used in this analysis was the following: (1) Energy, (2-19) MFCCs, (20) Energy delta, (21-38) MFCCs deltas (MFCC'), (39) Brightness (Br), (40) Bandwidth (Bw), (41) Zero Crossing Rate (ZCR), (42-51) Audio Spectrum Envelope, (52) Audio Spectrum Centroid, (53) Audio Spectrum Spread, (54-78) Audio Spectrum Flatness (54-78), and (79) L2-norm of total spectral energy envelope.

For the purpose of studying feature selection, we have initially chosen six audio events which we have considered to be very different: *applause*, *bird*, *dog (barking)*, *siren*, *telephone (analogic)*, and *water*. Table 3 shows the most important features for each concept, according to each method.

The first conclusion that can be retrieved from the results is that there is not a unique set of features that can be consider to be the best for all events. However, some features, such as MFCCs, are chosen for almost all events.

The number of selected features also differs widely, with more complex and diverse events needing more features than other events. For instance, for *applause* only 13 features were selected, whereas for *bird*, 33 were selected.

Concept	# CFS	CFS	InfoGain
applause	13	Br Bw Centroid Spread MFCC Envelope Flatness	Centroid Br+Bw+ZCR MFCC Envelope
bird	33	MFCC MFCC' Envelope L2-norm	L2-norm Envelope Br+ZCR Spread Envelope MFCC'
dog	22	Centroid Spread MFCC MFCC'	MFCC Centroid Br+ZCR MFCC' Envelope
siren	37	Centroid MFCC MFCC' Envelope Flatness L2-norm	Centroid Flatness Br+Bw MFCC Envelope ZCR L2-norm
telephone analogic	16	MFCC Flatness	Flatness Centroid Br+ZCR MFCC
water	22	Br Bw ZCR Centroid MFCC Envelope Flatness	Br+Bw+Zc Centroid MFCC Envelope Flatness

Table 3: Most important features according to the CFS and Information Gain measures.

The selected features also seem to depend on the spectral structure of the events. For events that have a flatter spectral structure, like *applause* or *water*, Brightness, Bandwidth, Zero Crossing Rate and MFFCs were considered the most important features. For events with a very pronounced harmonic nature, like *bird* or *siren*, features like Flatness, Envelope and the derivatives of the MFCCs become more useful.

¿From these results, we can conclude that the new added features can be very useful to audio event detection, specially for some specific events. However the initial set of features seems to be most useful for the generality of the audio events.

Since with PCA analysis we can reduce the size of the feature set, we considered that there is no disadvantage of using the same set of extended features for all events, and thus benefit from the extended information without increasing the complexity of the framework detection with a feature selection phase.

5. Data dimensionality reduction

Previous experiments [8] varying the analysis frame length have revealed the advantages of using large contexts in most cases, but they have also shown that smaller frames are more adequate

for computing differences between consecutive frames.

In an attempt to combine the best of both frame lengths, we computed the feature values for a central frame of 20 ms, and added the mean and standard deviation of 12 neighboring frames on each side, totaling 500 ms [13].

Since the number of features becomes prohibitively large for training SVM classifiers, this training was preceded by a PCA stage involving the above mentioned 79-feature set times 3. Features were normalized to guaranty that each feature had equal weight in the PCA stage. Using 150 principal components (PC) guaranteed a variance coverage of near 97.5% in the development set. Table 4 shows the results on the evaluation set, for the 6 selected concepts. Overall, the results are slightly better than the baseline results.

Concept	Test file	pr_p	pr_n	F-m	AP
applause	TS1	0.22	1.00	0.35	0.76
	TS2	0.17	1.00	0.29	0.52
bird	DOC1	1.00	0.89	0.08	1.00
	DOC2	0.03	0.80	0.01	0.00
	DOC3	1.00	0.96	0.56	0.95
dog	DOC4	0.53	0.92	0.65	0.74
	DOC5	0.19	0.98	0.29	0.20
siren	007-AViewToAKill	0.08	0.97	0.08	0.03
	DieHard4	0.00	0.81	0.00	0.00
	BN1	0.71	0.46	0.05	0.31
telephone analogic	TheMatrix	0.81	1.00	0.72	0.72
	TheAviator	0.77	1.00	0.58	0.67
water	DOC6	0.97	0.91	0.21	0.08

Table 4: SVM results for the evaluation set after feature dimensionality reduction to 150 PC (mean AP=0.459).

6. Perceptual Experiment

A subjective test was conducted with 23 subjects, using a set of 46 sound effect files. The purpose of this test was to evaluate the human confusability of audio events. Only the first 5s of each selected sound effect file was available for this test. Two files were selected as examples for each of 23 semantic concepts. All subjects (7 female, 16 male) had normal hearing, a University background, and their ages varied from 22 to 52. The number of times each file could be heard was not monitored. Given the fact that free answers were not specific enough (i.e. subjects did not know that there was a possibility of distinguishing between analog and digital telephone ringing), in the remaining tests, subjects were told examples of hypotheses, but could choose others if desired.

The tests confirmed that some of the most confusing tags are semantically very similar, such as traffic and car. In fact, one should revisit if a generic label such as traffic is good enough to encompass most terrestrial vehicles. Chickens were the hardest animals to detect, but one of the cat files was three times classified as baby crying. Overall, the hardest concept to classify was fire (most often classified as water).

The perceptual experiment provided useful cues for structuring our ontology of audio semantic concepts. We currently have over 100 concepts, although we have not yet built all detectors and will not likely have training material for all of them. The list includes both simple and *aggregated* events. An example of the latter is "animal", which corresponds to an event which will be triggered every time an event of a subclass of animals will be found.

7. Conclusions and future work

Our AED experiments have shown us that the performance of the classifiers in the sound effect corpus can be very different from the performance on the real data test set, where several audio events can coexist simultaneously and where recording conditions can be significantly different. Even so, the advantages of using an intrinsically labeled corpora, and the good results obtained in some audio events, justify this choice of training corpora. In an attempt to increase the robustness of our classifier, we are currently working on hierarchical clustering approaches, and adding harmonic related features.

We are also in the process of selecting the training/development material for animal sounds not previously modeled (Bear, Boar, Dolphins, Donkey, Elephant, Monkey, Moose_Elk, Rhinoceros, Seal, Shark, Snake-Hiss, Whale), and other sounds for which there is also enough material (Bite, Boat_Ship, Bullet, Drill, Elevator_Lift, Grinder, Ice, Morse-Code, Paper, Trolley, Wind). The remaining material includes musical instruments which are not the particular target of this project, sports (ball hit, jump, slide, etc.), human-voice sounds, and ambience sounds (Office, Bathroom, Kitchen, Market, Airport, Industry, etc.). The latter may be particularly hard to model given that they are characterized by a mixture of sounds.

8. References

[1] Meinedo, H., Audio Pre-Processing and Speech Recognition for Broadcast News, PhD Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2008.

[2] Xu, M. et al. "Creating audio keywords for event detection in soccer video", Proc. IEEE Int. Conf. on Multimedia and Expo, 2003.

[3] Cheng, W., Chu, W. and Wu, J., "Semantic context detection based on hierarchical audio models", Proc. 5th ACM SIGMM Int. Workshop on Multimedia information retrieval, 2003.

[4] Chu, W. et al., "A study of semantic context detection by using SVM and GMM approaches", Proc. IEEE Int. Conf. on Multimedia and Expo, 2004.

[5] Moncrieff, S. et al. "Detecting indexical signs in film audio for scene interpretation", Proc. IEEE Int. Conf. on Multimedia and Expo, 2001.

[6] Guo, G. and Li, S., "Content-based audio classification and retrieval by support vector machines", IEEE Trans. on Neural Networks, 14(1):209-215, 2003.

[7] Cai, R. et al. "A flexible framework for key audio events detection and auditory context inference", IEEE Trans. on Speech and Audio Processing, 2005.

[8] Trancoso, I. et al., "Training audio events detectors with a sound effects corpus", Proc. Interspeech 2008, Brisbane, Sept. 2008.

[9] Portêlo, J. et al., "Non-speech audio event detection, Proc. ICASSP'2009, Taiwan, April 2009.

[10] Chang, C. and Lin, C., "LIBSVM: a library for support vector machines", Manual, 2001. Online: http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[11] Hall, M., "Correlation-based Feature Selection for Machine Learning", PhD Thesis, The University of Waikato, 1999.

[12] Witten, I. and Frank, E., "Data mining: practical machine learning tools and techniques with Java implementations", ACM SIGMOD Record, 31(1), pp. 76–77, 2002.

[13] Temko, A., Nadeu, C. and Biel, J., "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07", Lecture Notes In Computer Science, Springer, volume 4625, pp. 354-363, 2008.