NON-SPEECH AUDIO EVENT DETECTION

José Portêlo[1], Miguel Bugalho[12], Isabel Trancoso[12], João Neto[12], Alberto Abad[1], António Serralheiro[13]

[1]INESC-ID Lisboa, Portugal
[2] IST, Lisboa, Portugal
[3] Military Academy, Portugal
Isabel.Trancoso@inesc-id.pt

ABSTRACT

Audio event detection is one of the tasks of the European project VIDIVIDEO. This paper focuses on the detection of non-speech events, and as such only searches for events in audio segments that have been previously classified as non-speech. 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 and HMM-based classifiers, using a 290-hour corpus of sound effects. Although we have only built detectors for 15 semantic concepts so far, the method seems easily portable to other concepts. The paper reports experiments with multiple features, different kernels and several analysis windows. Preliminary experiments on documentaries and films yielded promising results, despite the difficulties posed by the mixtures of audio events that characterize real sounds.

Index Terms— audio segmentation, event detection

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. Instead of carefully modeling each different semantic concept, the approach is to apply machine learning techniques to train many, possibly weaker detectors, describing different aspects of the audio-video content. The combination of many single class detectors will render a much richer basis for the semantics. The integration of cues derived from the audio signal is essential for many types of search concepts. Our role in the project is to contribute towards this integration with three different modules: audio segmentation, speech recognition, and detection of audio events. This paper concerns the last module.

Audio Events Detection (AED) 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. Optionally, there may be an intermediate stage of key audio effect detection, typically based on Hidden Markov Models (HMMs), that explores the time structure of the events and/or models interconnections between key audio effects (e.g. an explosion being preceded by a car crash).

The feature extraction process deals with different type of features, such as: total spectral power, sub-band power, brightness, bandwidth, MFCC (Mel-Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction), ZCR (Zero Crossing Rate), pitch frequency, 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. Many of these features are common to the audio segmentation and speech recognition modules. 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) [10], Gaussian mixture models (GMMs) [5] [4] [7] , Support Vector Machines (SVMs) [5] [6] [7] , and Bayesian Networks [2].

In this work we used HMMs and SVMs for building a one-against-all classifier for each semantic concept. This approach allows an easy extension to new semantic concepts, although better results could potentially be achieved by multiple-class classification.

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 [9]. The extended training corpus and the small test corpus of documentaries and movies will be described in section 2. The next section motivates our two-stage AED approach that first distinguishes between speech and non-speech audio events. Section 4 describes our multiple experiments with one-against-all detectors. Finally, section 5 presents the main conclusions and future plans.

2. CORPORA AND EVALUATION METRICS

The first corpus we considered for the task of audio events detection was a small pilot corpus of 422 sound effects files, totaling 6.8h, provided by B&G, one of the partners of the project. The choice of a sound effects corpus was made because it is intrinsically labeled, as each file typically contains a single type of sound. Since the initial results were quite promising, we moved on to a larger corpus of approximately 18700 files with an estimated total duration of 289.6h, also provided by B&G. The corpus includes enough training material for over 40 different audio events, but so far we have only considered 15. This initial list 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. However, many were recorded with a low 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 initial list of audio events. This *real life* corpus covers 13 of the 15 audio events.

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 work the detection performance (in every metric) is frame-based. Classification results in the test set are smoothed over time [9].

3. TWO-STAGE AED

Our initial experiments with the pilot sound effect corpus led us into adopting a two-stage approach for audio event detection. The first stage applies a speech/non-speech detector. This stage attempts to separate the events that are typically produced by the human speech production system (not only speech, but also laughing, crying, screaming, etc.), from the ones that are not related to human voice. In the second stage, separate classifiers attempt to detect either speech-related events or non-speech events, according to the initial classification. This paper addresses only the last category.

The original speech/non-speech (SNS_1) detector is based on an MLP (Multi-Layer Perceptron) trained with PLP features, extracted from a corpus of broadcast news. Although the performance of the classifier is very good for this domain [1], the type of non-speech events is quite limited (e.g. jingles). When tested in the sound effect corpus, the SNS sometimes detects speech in non-speech events, as shown in the fourth column of Table 1, which contains the duration of the detected speech segments.

This observation motivated the retraining of the detector including non-speech examples randomly selected from the large sound effects corpus (excluding the files that were used for training each audio event classifier). The results obtained with the new detector (SNS_2), given in the last column of Table 1, show an excellent false positive ratio (non-speech classified as speech), except for the Helicopter concept. Additionally, in a speech database, equivalent (speech) detection performance to the original SNS_1 detector was observed.

Audio event	#Files	Duration	SNS_1	SNS_2
Airplane_Jet	26	1210.2	31.6	0.0
Airplane_Propeller	58	2523.3	22.2	0.0
Birds	93	6339.8	106.9	0.0
Bus	34	2736.2	10.2	0.0
Cat_Meowing	42	1157.1	2.6	0.0
Crowd_Applause	30	1308.4	0.0	0.0
Dog_Barking	45	1860.5	35.7	0.0
Gun-Shot	110	2435.4	95.7	0.0
Helicopter	26	1298.5	56.8	28.8
Horse_Walking	85	3311.0	9.1	0.0
Sirens	47	1133.1	2.5	0.0
Telephone_Bell	17	562.3	6.2	0.0
Telephone_Digital	14	337.5	40.7	0.0
Traffic	32	4396.9	1.3	0.0
Water	72	6147.7	0.0	0.0
Total	762	36757.9	421.4	28.8

Table 1. List of audio events: number of files, total duration, and amount of data misclassified as speech (seconds).

4. ONE-AGAINST-ALL DETECTORS

With the objective of obtaining simple one-against-all detectors, we have built "concept-specific" and "world" models for the list of audio events. Our first experiments were carried out using the LIBSVM toolkit [3], for the 15 concepts. Then we performed in parallel experiments using the HMM toolkit from HTK [11] and feature dimensionality reduction techniques, for a restricted number of concepts.

4.1. SVM classifiers

The initial experiments were made with the purpose of evaluating the event detection results provided by different combinations of well-known features that will serve as a baseline for future comparisons. At this stage we only considered the use of PLP or MFCC (19 coefficients + energy + deltas) and 3 additional features: brightness, bandwidth and ZCR. The "world" model was build using between 92 and 96 files, 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)), but only the results for the RBF kernel are shown, as they

Event	MFCC+3	MFCC	PLP+3	PLP
AJ	**0.82**	0.77	0.78	0.76
AP	0.78	0.79	0.79	**0.81**
Bi	**0.90**	0.90	0.89	0.90
Bu	**0.90**	0.84	0.87	0.87
CM	0.75	0.71	0.80	**0.81**
CA	0.98	0.97	**0.99**	0.99
DB	0.95	**0.95**	0.90	0.90
GS	0.86	0.84	**0.87**	0.86
He	0.75	0.71	**0.82**	0.80
HW	0.92	0.92	**0.99**	0.95
Si	0.86	0.88	0.89	**0.90**
TB	0.80	0.85	0.84	**0.96**
TD	0.80	0.87	0.91	**0.91**
Tr	0.87	0.87	**0.91**	0.89
Wa	0.96	0.97	0.97	**0.97**

Table 2. SVM results for the development set (F-measure).

were overall better than the others. The results for these initial experiments on all considered audio events are presented in Table 2. The results obtained on the test set using the best combination of features on the development set are shown in Table 3. These results confirm that detecting audio events in real life data is much more challenging than the classification of isolated events. We expect that AED can benefit from incorporating time structure models and new features.

4.2. HMM classifiers: Modeling time structure

After the initial experiments with SVMs, we tried to take advantage of the periodic nature of some audio events. Although SVMs are a powerful machine learning tool, some other tools, like HMMs, are more suitable for modeling the time structure. Some of the 15 chosen audio events present a strong periodic nature, such as Airplanes, Helicopters and Sirens. We have chosen Sirens to test the HMM approach, due to their very distinct frequency characteristics. Left-to-right models with several number of states and Gaussian mixtures were trained to tune these parameters according to the development set results. MFCC features (12 coefficients + energy + deltas) of three different window lengths were used. In these experiments, the audio files have been down-sampled to 16kHz.

The results for the test set are shown in Table 4. These were obtained using the number of states and mixtures that yielded the best results on the development set. Even using a more limited feature set, the results for the 20ms window length show a small improvement over the previous SVM results (0.43 mean positive detection, compared with 0.29 for the SVMs). Only for the second file the results were worse.

4.3. Extended feature set

In the several experiments carried out throughout this work we could verify that the results of the SVM classifiers were highly dependent on the set of features. Since the Siren audio event has distinct frequency characteristics, we have explored an extended set of features that includes pitch. We have also tested a different method for representing feature variation, Shifted Delta Cepstrum (SDC) [8] (parameters: d=1,P=2,k=2). Because the pitch was extracted using 20ms windows and all the other features were extracted using 500ms windows, for every feature vector we included several pitch values. The total size of the extended feature vector is 52. Table 5 shows the results for the SVMs using PLPs (with deltas or SDC), the 3 additional features and pitch. The results were slightly worse compared to Table 3.

Event	Test file	pr_p	pr_n
AJ	TopGun	0.94	0.25
AP	TheAviator	0.66	0.90
Bi	DOC1	1.00	0.74
	DOC2	0.04	0.72
	DOC3	1.00	0.74
CA	TS1	0.29	0.98
	TS2	0.26	0.99
DB	DOC4	0.62	0.95
	DOC5	0.96	0.73
GS	TheMatrix	0.67	0.81
He	DieHard4	0.88	0.51
HW	007-AViewToAKill	0.24	0.35
Si	007-AViewToAKill	0.33	0.96
	DieHard4	0.49	0.94
	BN1	0.21	0.97
TB	TheMatrix	0.68	0.99
	TheAviator	0.76	0.99
TD	TheMatrix	0.00	1.00
	DieHard4	0.00	1.00
Tr	DieHard4	0.27	0.80
Wa	DOC6	0.45	0.94

Table 3. SVM results for the test set (pr_p and pr_n).

Test file	20ms		60ms		100ms	
	pr_p	pr_n	pr_p	pr_n	pr_p	pr_n
007	0.47	0.94	0.30	0.99	0.18	0.99
DieHard4	0.18	0.98	0.17	0.99	0.06	0.99
BN1	0.48	0.97	0.36	0.99	0.42	0.99
Total	0.43	0.97	0.32	0.99	0.29	0.99

Table 4. Results of training HMMs with several window lengths, for the test set (Sirens).

4.4. Data dimensionality reduction

The results obtained by adding the pitch feature have shown that increasing the number of features may decrease the performance of the SVMs. This motivated the use of PCA to perform feature dimensionality reduction on the Siren audio

Test file	PLP$_{deltas}$+3+pitch		PLP$_{SDC}$+3+pitch	
	pr_p	pr_n	pr_p	pr_n
007	0.24	0.97	0.32	0.97
DieHard4	0.28	0.96	0.18	0.98
BN1	0.41	0.97	0.38	0.99
Total	0.34	0.97	0.33	0.98

Table 5. SVM results with extended features (Sirens).

Test file	PCA	Variance coverage	PLP$_{deltas}$+3+pitch	
			pr_p	pr_n
007	2	85%	0.45	0.96
	3	90%	0.50	0.96
	10	95%	0.39	0.95
	20	99%	0.38	0.97
DieHard4	2	85%	0.32	0.99
	3	90%	0.29	0.98
	10	95%	0.14	0.97
	20	99%	0.14	0.98
BN1	2	85%	0.80	0.95
	3	90%	0.95	0.92
	10	95%	0.91	0.85
	20	99%	0.63	0.92
Total	2	85%	0.62	0.97
	3	90%	0.71	0.96
	10	95%	0.63	0.93
	20	99%	0.48	0.96

Table 6. SVM results with PCA features (Sirens).

event data. One of the advantages of PCA is to allow for a faster execution of the training process by reducing the number of features. Moreover, by combining the most discriminating features into a small set, the PCA removes unimportant data that can decrease the performance of machine learning algorithms such as SVMs. Table 6 shows the results using different numbers of PCA features. The principal components are calculated in the training set and their respective variance coverage rate is verified in the development set. The results show significant improvements relatively to the results using pitch, and are better than the initial SVMs results for the test set.

5. CONCLUSIONS AND FUTURE WORK

The initial experiments presented in this work allowed us to conclude 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. We are currently working towards reducing the differences between the training/development and test data by using normalization techniques, and we are also testing agglomerative clustering approaches. We observed that HMMs are a promising method for our AED task that justifies further tests. The use of feature dimensionality reduction methods is also worth pursuing, particularly when dealing with several features that may influence differently the detection of acoustic events.

6. REFERENCES

[1] Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I., and Neto J., "A Prototype System for Selective Dissemination of Broadcast News in European Portuguese", EURASIP J. on Adv. in Signal Processing, Hindawi Publishing Corporation, vol. 2007, n. 37507, May 2007.

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

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

[4] 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, pages 109-115, 2003.

[5] 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.

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

[8] Torres-Carrasquillo, P. A. et al. "Approach to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", Proc. ICSLP 2002, Denver, September 2002.

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

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

[11] Young, S. et al. "HTK - Hidden Markov Model Toolkit", Manual, 2006. Online: http://htk.eng.cam.ac.uk/