AUDIO CONTRIBUTIONS TO SEMANTIC VIDEO SEARCH

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

[1]INESC-ID Lisboa [2] IST/UTL, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal

ABSTRACT

This paper summarizes the contributions to semantic video search that can be derived from the audio signal. Because of space restrictions, the emphasis will be on non-linguistic cues. The paper thus covers what is generally known as audio segmentation, as well as audio event detection. Using machine learning approaches, we have built detectors for over 50 semantic audio concepts.

Index Terms— Audio Segmentation, Audio Event Detection

1. INTRODUCTION

The framework for this work is the European project VIDI-VIDEO, 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 main modules: audio segmentation, audio event detection, and speech recognition. Because of space restrictions, the paper concerns only the first two modules. Other modules which could be potentially very interesting to semantic video search such as topic classification and language identification will not be covered here, for the same reason.

Three application scenarios are considered in the project: broadcast news (BN), cultural heritage, and surveillance. The latter very often does not include audio, which justifies our major efforts with BN shows and cultural documentaries.

Audio segmentation can mean many different things. In this paper, we restrict its meaning to the type of segmentation that can be performed on the audio signal alone, without taking into account its linguistic contents. This type of segmentation can be done in several tasks. Acoustic Change Detection (ACD) is the task responsible for the detection of audio locations where speakers or background conditions have changed. Speech/Non-speech (SNS) classification is responsible for determining if the audio contains speech or not. Gender Detection (GD) distinguishes between male and female gender speakers, but an age-directed segmentation can be also useful for detecting children voices, for instance. Background Conditions (BC) classification indicates whether the background is clean, noisy or musical. Speaker Clustering (SC) identifies all the speech segments produced by the same speaker. Speaker Identification (SID) is the task of detecting the identity of certain often recurring speakers like news anchors or very important personalities. More recently, the term speaker diarization (SD) became used to mean segmentation into speaker-homogeneous regions, answering the question "Who spoke when?".

Whereas audio segmentation has been around for quite a while, the area of audio event detection (AED) is much more recent. 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, 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, many of them common to the ones found in audio segmentation or speech recognition modules, or the MPEG-7 descriptors. Due to the potentially large amount of features, which can lead to slow convergence of the classification algorithms, the use of feature dimensionality reduction techniques like Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) is also very common. In the inference process, various machine learning methods are used [1] [3] [6],such as rule-based approaches, Gaussian mixture models (GMMs), Support Vector Machines (SVMs), and Bayesian Networks.

This paper includes two main Sections dedicated to audio segmentation and audio event detection, respectively. These modules contribute to the detection of a wide range of semantic concepts, whose ontology is summarized in Section 4.

2. AUDIO SEGMENTATION

The audio segmentation module [5] includes six separate components: one for Acoustic Change Detection, four components for classification (Speech/Non-speech, Background, Gender and Speaker Identification) and one for Speaker Clustering. These components are mostly model-based, making

extensive use of feed-forward fully connected MLPs (Multi-Layer Perceptrons) trained with the back-propagation algorithm. All the classifiers share a similar architecture: an MLP with 9 input context frames of 26 coefficients (12th order Perceptual Linear Prediction (PLP) plus energy and deltas), two hidden layers with 250 sigmoidal units each and the appropriate number of softmax output units (one for each class), which can be viewed as giving a probabilistic estimate of the input frame belonging to that class.

Despite the Acoustic Change Detection and Speech/Non-speech blocks being conceptually different, they were implemented simultaneously in the SNS component, considering that a speaker turn is most often preceeded by a small non-speech segment. The output of the SNS MLP classifier is smoothed using a median filter, and processed by a finite-state machine, involving confidence and duration thresholds.

When a speaker change is detected, the first t_{sum} frames of that segment are used to calculate gender, background conditions, and speaker identification (anchors) classifications. Each classifier computes the decision with the highest average probability over all the t_{sum} frames.

The Speaker Clustering component which uses an online leader-follower strategy tries to group all segments uttered by the same speaker. The first t_{sum} frames (at most) of a new segment are compared with all the same gender clusters found so far. Two SC components are used in parallel (one for each gender). A new speech segment is merged with the cluster with the lowest distance, provided if it falls below a predefined threshold. The distance measure for merging clusters is a modified version of the Bayesian Information Criteria.

Our latest addition to the audio segmentation module is a telephone bandwidth detector. Given the lack of a large manually labeled corpus, a bootstrapping approach has been adopted in which a simple LDA classifier has been trained with a small amount of manually labeled data in order to generate automatic transcriptions for the posterior development of a binary MLP classifier. The adopted feature set consisted of 15 logarithmic filter bank energies extracted at a frame rate of 20 ms with a time shift of 10 ms, and corresponding deltas.

2.1. Results

	CER / DER
Speech/Non-Speech	4.7
Gender	2.4
Speaker Clustering (anchors)	4.1
Speaker Clustering (all)	26.0

Table 1. *Audio segmentation evaluation results.*

Table 1 summarizes the evaluation results conducted in a test set of six 1-hour long BN shows, collected during 2007. The results are shown in terms of frame Classification Error Rate (CER) for SNS and GD blocks and by Diarization Error Rate (DER) for SC. They correspond to a t_{sum} delay of 4s. The results can be considered comparable to state of the art for the SNS and GD components. The BC manual labels lacked consistency. The speaker clustering performance for news anchors shows very good results due to the SID models. For the other speakers the results are not so good. In part these DER results can be attributed to the long duration of the BN shows which have an average of 64 different speakers per news show, and also to the very large percentage of speech with loud background noise, mainly from street interviews.

The telephone bandwidth classifier was not evaluated in this data set, which did not include telephone data labels. The rate of correct classified frames in the validation data set obtained by the LDA classifier was 99.8%. In other BN test sets, the rate achieved by the MLP was lower, which we also attributed to the high variability in the training data.

The whole audio segmentation block operates in 0.014 xRT which is very adequate for its potential application to on-line subtitling of TV shows.

3. AUDIO EVENT DETECTION

The lack of corpora manually labeled in terms of audio events motivated the adoption of a very large sound effect corpus for training, given that it is intrinsically labeled, as each file typically contains a single type of sound. The corpus includes approximately 18,700 files with an estimated total duration of 289.6h, and was provided by one of the partners in the project (B&G). It includes enough training material for nearly 50 different audio events, not related to human voice, i.e., it excludes events such as laughing, crying or yawning. In fact, the application of the SNS to this set of files, yields very few cases where speech is erroneously detected.

The list of events considered is shown in Table 2, 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).

This corpus was used to train one-against-all detectors for each concept by building "concept-specific" and "world" models. Our initial set of detectors was SVM-based, and the experiments were made using the LIBSVM toolkit [2]. Preliminary experiments [8] compared the performance of a limited set of features: PLP or MFCC (Mel-Frequency Cepstral Coefficients) coefficients (19 + energy + deltas), ZCR (Zero Crossing Rate), brightness, and bandwidth. The latter are, respectively, the first and second order statistics of the spectrogram, and they roughly measure the timbre quality.

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 func-

tion (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 2. 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	Abb.	#Files	Duration	F-m
airplane jet	jet	26	1210.2	0.823
airplane propeller	pr	58	2523.3	0.811
bell electric	beE	34	704.2	0.723
bell mechanic	beM	117	4669.5	0.845
big cat	biC	91	3038.5	0.885
bird	bir	93	6339.8	0.905
bus	bus	34	2736.2	0.900
buzzer	buz	23	503.3	0.722
car	car	97	3722.2	0.946
cat	cat	40	1110.6	0.812
chicken	chi	16	434.6	0.863
cow	cow	24	692.8	0.705
crowd applause	app	30	1308.4	0.993
digital beep	diB	38	970.1	0.879
dog	dog	45	1860.5	0.954
door	doo	113	1059.8	0.502
explosion	exp	43	672.1	0.888
fire	fir	53	4706.6	0.939
firework	fiW	22	692.9	0.456
frog toad	fro	50	2775.0	0.924
glass	gla	58	977.3	0.872
gunshot heavy	guH	37	972.4	0.926
gunshot light	guL	110	2435.4	0.872
hammer	ham	45	1469.9	0.671
helicopter	hel	26	1298.5	0.822
horn	hoV	80	1089.5	0.944
horse	hor	85	3311.0	0.950
insect buzz	inB	28	1823.5	0.976
insect chirp	inC	33	3267.2	0.987
motorcycle	mot	132	7784.0	0.941
pig	pig	33	1490.0	0.743
rattlesnake	rat	32	773.6	0.994
saw electric	saE	38	1290.7	0.883
saw manual	saM	24	887.8	0.625
sheep	she	33	1602.8	0.912
siren	sir	47	1133.1	0.901
telephone analogic	be	17	562.3	0.966
telephone digital	dig	14	337.5	0.914
thunder	thu	52	1941.2	0.980
traffic	trf	32	4396.9	0.911
train	tra	82	4895.5	0.889
typing	typ	32	2434.2	0.945
walking hard	waH	93	4014.3	0.930
walking soft	waS	86	4421.7	0.934
water	wat	72	6147.7	0.976
whistle	whi	46	601.3	0.836
wolf howling	woH	31	1006.8	0.935

Table 2. Number and total duration of files for each of the 47 concepts, together with F-measure results for the development set.

3.1. Results

In order to test the detectors in a *real life* situation, we manually labeled a number of movies, documentaries (DOC), talk shows (TS) and broadcast news (BN) in terms of 13 audio events. The experiments with the test corpus were 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 movies. The detection performance is frame-based, but classification results in the test set are smoothed over time.

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. The worse performance can often 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
jet	TopGun	0.94	0.25
pro	TheAviator	0.66	0.90
bir	DOC1	1.00	0.74
	DOC2	0.04	0.72
	DOC3	1.00	0.74
app	TS1	0.29	0.98
	TS2	0.26	0.99
dog	DOC4	0.62	0.95
	DOC5	0.96	0.73
gun	TheMatrix	0.67	0.81
hel	DieHard4	0.88	0.51
hor	007-AViewToAKill	0.24	0.35
sir	007-AViewToAKill	0.33	0.96
	DieHard4	0.49	0.94
	BN1	0.21	0.97
beC	TheMatrix	0.68	0.99
	TheAviator	0.76	0.99
dig	TheMatrix	0.00	1.00
	DieHard4	0.00	1.00
trf	DieHard4	0.27	0.80
wat	DOC6	0.45	0.94

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

3.2. Hierarchical clustering

Even if the sound effect corpus provides sound files with precise titles, there are files that do not correspond to their associated semantic concept. These outliers are not suitable to train detectors or classifiers. For example, among the sound samples of the Bird semantic concept, there is a woodpecker pecking wood, with no singing. To avoid the very morose task of listening to the audio samples of all the 47 concepts, unsupervised clustering, and agglomerative clustering in particular, has been investigated. In the literature, the sound detection and identification system named SOLAR [4], or the system developed in the CUIDADO project, for musical instruments [7], use a hierarchical architecture. Large categories of sounds are considered first, for example, sustained and non-sustained sounds, and then more precise classes are detected. In this study, agglomerative clustering aimed also at defining without knowledge the sound classes that could be used in a hierarchical detection system.

Agglomerative clustering or hierarchical clustering (HC) iteratively merges patterns into bigger clusters. Data items within each cluster are more closely related to one another than to items assigned to other clusters. Therefore, the last items to be merged are potential outliers, since they involve

the biggest distances with the other patterns. HC has been performed within each semantic concept at pattern level. For the Bird concept for example, the pecking Woodpecker was among the last samples to be clustered.

Clusters at concept level showed perceptually or semantically similar concepts. Figure 1 shows the dendrogram for all the 47 concepts, for the positive examples of the training set. The features used were PLP+deltas+3; only means and variances were used. The most similar concepts that have been found are Bus, and Traffic. Another cluster involves Door and Hammering concepts. More generally, a rough separation between "non-pitched" sounds (concepts: bus, traffic, car, etc.) and "pitched" sounds (concepts: bird, dog, chicken, etc.) has been identified. In a hierarchical AED system, a pitched/non-pitched classifier could be trained with these two groups.

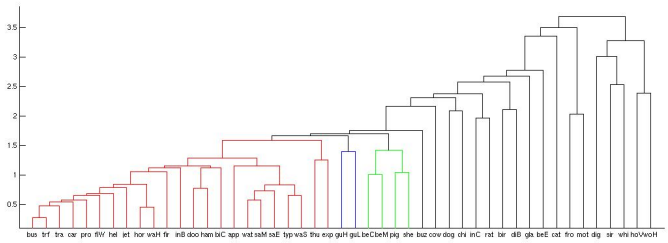

Fig. 1. Dendrogram for the 47 concepts.

4. ONTOLOGY OF AUDIO SEMANTIC CONCEPTS

The work described above contributed to building a list of concepts which can be derived from the audio signal, without taking into account its linguistic contents. 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. Most of these concepts are derived from the AED module and less than 10 from the audio segmentation module (e.g. female voice, music, low-quality audio). Some concepts also involve more or less sophisticated strategies for combining information from the audio segmentation module over a complete multimedia document, such as "dialogue". Music genre classification, although not contemplated in the project, also contributes to this list of potential concepts (e.g. violin music, piano music, drums, choir, orchestra, jazz, classic music, etc.).

The VIDIVIDEO partners responsible for the integration of audio and video cues derive the audio cues from XML files, whose format follows the MPEG-7 standard. The audio segmentation may either produce a single file with information about all the related concepts, or separate files for each concept (e.g. *female_voice.xml*). Each AED detector produces an XML file with a similar structure (e.g. *dog_barking.xml*). The corresponding confidence level is also included.

5. CONCLUSIONS AND FUTURE WORK

This paper summarized our work in terms of audio segmentation and audio event detection. Our current work in terms of audio segmentation spans several directions in parallel. We are working on a 3-class gender detection module (male, female and child), and on an improved speech/music classification module. We are also trying to improve the ACD performance in cases where speakers interrupt/overlap each other, or where recordings from different speakers are artificially concatenated without inserting pauses. In what concerns speaker clustering, the addition of other features such as MSG (Modulation Spectrogram) also yields promising results. In terms of AED, we expect to benefit from incorporating time structure models, new features and dimensionality reduction techniques, and to follow up on the promising hierarchical clustering approach. The integration with video cues also leads to a wide range of interesting research topics.

6. REFERENCES

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

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

[3] 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.

[4] Hoiem. D, Yan, K., and Sukthankar, R., "SOLAR: sound object localization and retrieval in complex audio environments", Proc. ICASSP 2005, Philadephia, USA, 2005.

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

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

[7] Peeters, G., and Rodet X., "Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments", Proc. DAFX03, London, UK, 2003.

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