Study and application of acoustic information for the detection of harmful content, and fusion with visual information.

Theodoros Giannakopoulos [⋆]

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

Abstract. This thesis aims at investigating and developing techniques for content-based segmentation and classification of multimedia files, based on audio information. Emphasis has been given to analyzing the content of films based on audio information. In addition, part of the thesis is focused on the detection of audio classes related to *violent* content (e.g., gunshots, screams, etc).

1 Introduction

During the last decades, with the advances in the Word Wide Web and in the storage technology, an enormous increase of the available multimedia files has occurred. This explosion in the amount of the multimedia files being stored, transmitted and accessed has led to several research efforts focused on automatically and semantically analyzing the respective information. This is the task of ***content-based multimedia analysis***. This thesis has focused on developing methods for content-based analysis of multimedia files, based on the audio information. Focus has been given on issues such as: speech-music discrimination, general audio segmentation, multiclass classification and speech-based emotion recognition.

A very important issue related to the increase of the multimedia files (especially for those available through the World Wide Web) is that they are easily accessible by large portions of the population, with limited central control. It is therefore obvious that the need of protection of sensitive social groups (e.g., children) is imperative. Towards this end, several methods for violence detection in video files have been proposed, though in most of these methods, no audio information is used, or such use is limited to simple, energy-based features. However, the **audio** channel of a movie (or any video file) is very informative with respect to the content-based classification, especially when violence is the main target, because most violence-related content classes can be more easily detected through the usage of the audio data. For example, it is difficult (and most of the times impossible) to detect a gunshot in a video file by using only visual cues, but using the audio signal this task is much easier. Therefore, an important

[⋆] Dissertation Advisor: Sergios Theodoridis, Professor part of the present thesis is to **detect violent content** in videos, using audio classification techniques.

Finally, an emphasis in this thesis was to develop **annotated databases** which were used, both for training and evaluating the several segmentation and classification methods. To this end, two different types of databases have been used. One from radio recordings, which was used in the speech-music discrimination task, and another one from a number of movies.

2 Features

Feature extraction is a very important stage for audio analysis and processing tasks. In this thesis, several features have been studied and used for the proposed classification and segmentation methods. The choice of the specific features is the result of extensive experimentation and conclusions that stem from the physical meaning of the audio signals. Therefore, among the theoretical description of the audio features, some examples of differentiation of those features for different audio classes have been presented. All features have been extracted using a short-term processing technique ([1]), which results in a feature sequence for each audio signal, while for each segment (mid-term window) a statistic is computed over that feature sequence (e.g. the standard deviation). In general, the following features have been used in the thesis by the classification and segmentation algorithms:

- **Signal Energy:** Experiments have shown that the variations of the energy sequences in speech signals are higher than in music signals.
- **Zero Crossing Rate (ZCR):** This feature can be used for discriminating noisy environmental sounds, e.g., rain. Furthermore, in speech signals, the $\frac{\sigma^2}{\mu}$ ratio of the ZCR sequence is high, since speech contains unvoiced (noisy) and voiced parts and therefore the ZCR values have abrupt changes. On the other hand, music, being largely tonal in nature, does not show abrupt changes of the ZCR.
- **Energy Entropy:** This is a measure of abrupt changes in the energy level of an audio signal. Energy entropy can be used for discrimination of abrupt energy changes, e.g. gunshots, abrupt environmental sounds, etc.. In order to detect abrupt sounds of violent content (e.g., gunshots, explosions and fights) the feature energy entropy has also been used in a number of publications. Though, since this feature only contains energy-related signal information, it has been used in combination with other audio features ([2], [3], [4]), or in combination with visual cues ([5]).
- **Spectral centroid:** This is the center of "gravity" of a signal's spectrum. Experiments have indicated that the sequence of spectral centroid is highly variated for speech segments, while for sounds of human screams the deviation of the spectral centroid is significantly low.
- **Spectral Rolloff:** This is the frequency below which certain percentage of the magnitude distribution of the spectrum is concentrated. Experiments

have shown that for a large majority of the speech segments the mean value of this feature is around 0.50. Furthermore, for the “gunshots” segments the mean value has been found to be significantly higher.

- **Spectral Entropy:** This feature is computed by dividing the spectrum of the short-term frame into sub-bands (bins). The normalized energy each sub-band, is then computed and the entropy of that sequence of normalized energies is finally extracted. As part of this thesis, a variant of the spectral entropy called “chromatic entropy” has been used in [6] and [7] in order to discriminate in an efficient way speech from music.
- **Chroma-based features:** Chroma vector is a 12-element representation of the spectral energy, which encodes and represents harmonic relationships within a particular music signal. When the chroma vector is computed for each frame of the audio segment, a sequence of chroma vectors is formed, which is known as the **chromagram**. In this work, two features, based on the chromagram, have been proposed, in order to discriminate between speech and music. Though, experiments have shown that these features have a high discrimination accuracy for other audio classes, e.g. for gunshots.

3 Speech-music discrimination

Speech/Music discrimination (SM) refers to the problem of segmenting an audio stream and labelling (i.e., classifying) each segment as either speech or music. In this thesis, a multi-stage method for this task has been proposed, based on dynamic programming and bayesian networks. The proposed system proposed is based on a three-stage philosophy:

Chromatic Entropy Segmenter (*CES*): A computationally efficient algorithm is first employed as a preprocessing stage. It is based on a region growing technique that bears its origins in the field of image segmentation and operates on a single feature, which is a variant of the spectral entropy. A useful property of this very simple algorithm is that it can easily be *tuned to maximize speech or music precision* at the expense of leaving certain parts of the audio recording unclassified. To exploit this property, the algorithm is **applied twice** on the original recording: in the first pass, it is tuned to detect music segments with a high precision rate and during the second pass to yield speech segments with a high precision rate. After the application of this scheme, an amount of data is **left unclassified**. However, the *precision rate* of those which have been classified is *over* 98%.

Dynamic Programming Based Segmenter: At a second stage, a more sophisticated and computationally demanding algorithm is applied on the regions left unclassified. Each one of these regions is first split into a number of short-term frames by means of a short-term processing window and five features are extracted per frame. Speech/music discrimination is then treated as a **maximization task**. In other words, the method processes the feature sequence in order to *group features together and form the sequence of segments* and the *respective class labels* (i.e., speech/music) that *maximizes* the product of posterior probabilities, given the data that contribute to each one of the segments. In order to estimate the required posterior probabilities, a Bayesian Network Combiner is trained and used. Since an exhaustive approach to this solution is unrealistic, we resort to **dynamic programming** to solve this maximization task.

Post-process: In the final stage, a boundary correction algorithm is applied on the previously obtained discrimination results, in order to improve the overall system's accuracy. This algorithm locates the boundary that maximizes the product of the probabilities (generated by a Bayesian Network), so that the left and right segments (speech and music or vise versa) are correctly classified.

The overall architecture has been tested on a manually segmented and labelled dataset that consists of 9 hours of uninterrupted audio recordings from various radio broadcast genres (e.g., classical, pop-rock, news, etc.). The average confusion matrix and respective accuracy (over all genres) for each method is displayed in Table 1.

Table 1. Average confusion matrix (over all genres) and respective overall accuracies (A) per method.

	CES		**DPBS**		**Overall**		**Overall2**	
	M	**S**	**M**	**S**	**M**	**S**	**M**	**S**
M	69.09	1.59	69.24	1.44	69.34	1.34	69.53	1.15
S	6.74	22.58	4.18	25.14	3.51	25.80	3.17	26.15
	A: 91.67		**Ov. A: 94.38**		**A: 95.15**		**A: 95.68**	

A conclusion drawn from these results is that when CES and DPBS are used independently, as standalone techniques, DPBS offers an enhanced performance compared to CES for most of the genres. Furthermore, it is obvious that combining these two techniques (CES as a preprocessing stage), it only results to an extra gain of the order of 1% with respect to the best individual performance. The obvious question is whether this extra gain really justifies the combination of CES and DPBS. However, the main reason of using CES as a preprocessing stage was primarily of computational nature (the CES algorithm is computationally light). Furthermore, experimentation revealed that on the average, 54% of the audio stream is pre-segmented and classified using the CES algorithm (the rest of the data is segmented with the DPBS). Thus, besides a 1% performance gain, employing the CES as a pre-segmentation step leads to a significant reduction in the overall execution time. Finally, the results show that the post-processing step leads to an extra 0.5% performance improvement at only a little extra computational cost.

4 Music tracking in movies

Music tracking in audio streams can be defined as the problem of *locating the parts of an audio stream that contain music, possibly overlapping with other types of audio.* The problem of music tracking in audio streams has recently attracted a lot of attention, mainly in the context of audio content characterization applications. In the general case, music tracking is a hard task, because music is frequently mixed with other audio types. This is more apparent in the case of audio streams from **movies**, due to the diversity of sound sources involved in a film's soundtrack. *In the present work, no assumptions concerning the types of audio to be encountered in the stream have been made. This was the most important challenge of this task, along with the need for a computationally efficient method.*

At a first step, the audio signal is mid-term processed with a moving window technique. The goal is to extract four features per mid-term window. Each feature is a statistic, computed over a sequence of short-term features comprising the mid-term window. In particular, the following four features / statistics have been used: 2 chroma-based features, maximum energy entropy and non-zero pitch ratio ([8]).

After the feature extraction stage, a probabilistic soft output is calculated, i.e., a measure of confidence that the input sample has been extracted from a music segment. At this point, statistical independence has been assumed, in order to estimate the individual probabilities. The sequence of soft decisions for all mid-term segments is then processed by means of a median window to remove spurious values. A hard threshold is then applied in order to detect music segments. After extensive experimentation, the recommended threshold value was chosen to be equal to 0.1.

In order to evaluate the performance of the proposed music tracking system, audio streams from eight movies have been manually annotated (2.5 hours total duration). The performance (classification and detection) of the proposed algorithm is presented, in detail, in Table 2.

Table 2. Classification and detection performance for threshold value $T = 0.1$.

	Precision	**Recall**	**F1 Measure**
Classification	89%	83%	86%
Detection	91%	90%	90.5%

The results indicate that the method is accurate when music is in the background of other audio events, while experiments have shown that the computational complexity is kept quite low (around 10% of the recording time). The proposed music tracker can be used in an overall system for multi-class audio classification, as a pre-processing stage.

5 Audio segmentation

The purpose of audio segmentation is to *locate changes* in the content of the audio signals; in other words, to detect changes among acoustically homogenous audio regions. It is an important preprocessing step in any audio characterization system. Music information retrieval, video segmentation and audio characterization in security surveillance systems are some notable applications of high current interest. In such systems, besides accuracy, computational time is also of paramount importance, especially when a real-time or almost a real-time operation is desirable.

In general, audio segmentation approaches can be categorized into supervised and unsupervised techniques. Supervised approaches, e.g., [9], use a group of *a-priori known* audio classes and audio segmentation is performed via a classification task, by assigning audio frames in the respective classes. Unsupervised techniques treat audio segmentation as a hypothesis test by detecting changes in the audio signal, given a specific observation sequence ([10], [11]). Another differentiation between audio segmentation methods is that, depending on the task, the definition of homogeneity may vary.

In this thesis, the supervised approach rationale is followed, albeit using a completely different viewpoint, compared to previously developed techniques. Since all it is required is to detect content "changes" in the audio stream, we focus on this task *directly*, instead of solving another problem first (i.e., a classification task) and trying to infer our desired goal from it. Using this path, no a-priori assumption on the number of audio classes is required, which in a general audio stream cannot be easily determined. The problem of detecting the limits of homogenous audio segments is treated as a **binary classification** task. Each audio frame is classified as **"segment limit" vs "non-segment limit"**. For each frame, a spectrogram is computed and eight feature values are extracted from respective frequency bands. Final decisions are taken based on a classifier combination scheme. The algorithm has very low complexity with almost real time performance.

The algorithm has been evaluated on real audio streams from movies and it achieves 85% accuracy rate. Moreover, it introduces a general framework to audio segmentation, which does not depend explicitly on the number of audio classes.

6 Multi-class audio classification

A multi-class classification algorithm for audio segments recorded from **movies** has been presented, focusing on the detection of **violent** content, for protecting sensitive social groups (e.g. children). The task of detecting violence is difficult, since the definition of violence itself is ambiguous. In video data, most violent scenes are characterized by specific audio signals (e.g. screams and gunshots).

The literature related to violence detection is limited and, in most of the cases, it examines only visual features ([12], [13]). In [5] the audio signal is used as additional information to visual data. In particular, a single audio feature, namely the energy entropy, is used in order to detect abrupt changes in the audio signal, which, in general, may characterize violent sounds. In [14], a film classification method is proposed that is mainly based in visual cues, since the only audio feature adopted is the signal's energy. A more detailed examination of the audio features for discriminating between violent and non-violent sounds was presented in [4]. In particular, seven audio features, both from the time and frequency domain, have been used, while the binary classification task (violent and non violent) was accomplished via the usage of Support Vector Machines.

This thesis focuses on detecting violence in audio signals but also on giving a more detailed characterization of the content of those signals. Therefore, facing the problem as a binary classification task (violent/non-violent) would not be adequate. In addition, such a treatment of the problem would be insufficient in terms of classification accuracy. Seven classes (3 violent and 5 non-violent) have been defined, motivated by the nature of the audio signals met in most movies. The non-violent classes are: *Music*, *Speech*, *Others1*, and *Others2*. The later two non-violent classes are environmental sounds met in movies. These sounds have been divided into two sub-categories according to some general audio characteristics. In particular, "Others1" contains environmental sounds of low energy and almost stable signal level (e.g. background noise). "Others2" contains environmental sounds with abrupt signal changes, e.g. a door closing etc. As far as the *violent*-related content is concerned , the following classes have been defined: *Shots*, *Fights* (beatings) and *Screams*.

12 audio features are extracted for each segment on a short-term basis, i.e., each segment is broken into a sequence of non-overlapping short-term windows (frames), and for each frame a feature value is calculated. This process leads to 12 feature sequences. In the sequel, a statistic is calculated for each sequence, leading to a 12-D feature vector for each audio segment. In order to achieve multi-class classification, the "One-vs-All" (OVA) classification scheme has been adopted, which is based on decomposing the K-class classification problem into K binary sub-problems. In the current work, we have chosen to use Bayesian Networks (BNs) for building those binary classifiers. At a first step, the 12 feature values $v_i, i = 1 \ldots 12$ are grouped into three 4D separate feature vectors. In the sequel, for each one of the 7 binary sub-problems, three k-Nearest Neighbor classifiers are trained on the respective feature space. This process leads to 3 binary decisions for each binary classification problem. In order to classify the input sample to a specific class, the kNN binary decisions of *each subproblem* are fed as input to a separate BN, which **produces a probabilistic measure for each class**. After the 7 probabilities are calculated for all binary subproblems, the input sample is classified to the class with the largest probability.

Seven datasets D_i, $i = 1 \ldots 7$ consisting of 200 minutes of movie recordings have been compiled for training and evaluation reasons. Almost 5000 of audio samples have been extracted and manually labelled. The duration of those audio segments varies from 0.5 to 10 seconds. The data was collected from more than 30 films, covering a wide range of genres. In order to test the overall classification system, hold-out validation has been used. The normalized average confusion matrix (C) is presented in Table 3. The overall classification accuracy (i.e. the percentage of the data that were correctly classified) of the proposed method is 69.1%. Though this is a high performance rate according to the nature of the problem, one may prefer to use the proposed classification scheme as a binary classifier. The violence recall was found equal to 84.8% and the violence precision equal to 83.2%. This means that the overall **binary** classification accuracy was almost 84%.

Table 3. Average Confusion Matrix

	Classified						
True ↓	Mu	Sp	Ot1	Ot2	Sh	Fi	Sc
music	68.22	2.36	13.60	1.76	3.27	3.83	6.95
speech	1.66	81.96	6.38	4.75	0.23	2.08	2.95
others1	4.59	1.90	70.24	11.20	5.44	2.52	4.11
others2	2.00	3.15	15.21	59.83	10.30	8.57	0.94
shots	1.26	0.19	3.00	6.66	79.10	9.68	0.11
fights	1.70	2.23	0.89	11.81	26.38	52.29	4.71
screams	9.18	3.44	4.00	1.29	2.20	7.86	72.04

7 Emotion recognition in speech from movies

Besides extracting information regarding events, structures (e.g., scenes, shots) or genres, a substantial research effort of several multimedia characterization methods has focused on recognizing the **affective** content of multimedia material, i.e., the **emotions** that underlie the audio-visual information ([15], [16], [17]). In this thesis, emphasis has been given on affective content that can be retrieved from the speech information of movies. This approach can also help in detecting oral violence in movies, based on the emotional recognition results. This is very important, since oral violence is quite often present in films and it may sometimes be more harmful for children than physical violence. Note that emotion recognition in movies is a difficult task, since both audio and visual channels are more complicated in movies than in similar studio-acted databases.

A computationally efficient algorithm for tracking speech in audio streams from movies is firstly presented. The proposed algorithm achieves a precision rate of 95%. In order to find the emotional state of the detected speech segments from movies, we have proposed a 2-D representation (Arousal-Valence). To investigate whether the Arousal-Valence representation (Emotion Wheel) is appropriate for speech signals, several humans have manually annotated speech segments using this representation. If the Emotion Wheel is a good representation, then the differences in annotation by separate humans should be relatively small and the respective perceptions should be, on average, in good agreement. An extensive experimentation has led to the final selection of certain audio features and then the regression problem of mapping the feature space to the emotional plane is defined. Three regression schemes are evaluated using the annotated data, and the performance errors are compared to the error of the human annotation. An overall scheme for emotion recognition of large audio streams is proposed, that combines: a) the novel speech tracking algorithm, b) a segmentation algorithm that detects homogenous speech segments and c) the proposed method for emotion recognition of these speech segments.

The major contribution of this chapter is the novel dimensional approach for emotion recognition of speech segments from movies. Besides testing the emotion recognition methods, we have also focused on evaluating the emotional representation itself. The Emotion Wheel has been found to be a good representation of the affective content of speech segments, since the corresponding manual annotations performed by several humans were in good agreement. Moreover, experiments have shown that the regression performance of all three methods was high, since the error was comparable to the average error of the manual (human) annotations. This means that the proposed audio features can be successfully mapped to the emotion plane. Finally, we have demonstrated how to extract emotional information from uninterrupted audio streams from movies.

8 Conclusions

This thesis investigated methods for content-based characterization of multimedia data based on the audio information. Emphasis has been given on two different types of content: audio recorded from radio broadcasts and audio recorded from movies. In the first case we propose a novel speech - music discrimination algorithm, while in the second several audio analysis algorithms (emotion recognition, multi-class classification, etc) have been proposed.

A challenging and promising research issue for the future, is the development of a **joint segmentation - classification** method for the multi-class problem. Furthermore, the Bayesian Network classifier could be expanded with more nodes. Those nodes could represent other types of individual classifiers (e.g. Support Vector Machines). Apart from implementing other types of individual classifiers, some individual decisions based on **other types of media** (e.g. image, text) can be added. Another promising future issue is to enhance the performance of the overall movie characterization system, by implementing **music classification** algorithms. Often music tracks in movies reveal particular semantic meanings, e.g., the emotional tension of a scene in a horror film. Apart from that, the musical genre itself may provide us with important information about the content of a movie. Finally, a challenging task will be to develop methods for movie (or generally video) search, based on audio analysis. The audio class-specific probabilistic measures, the speech emotional representation, along with other types of audio analysis (e.g. music recognition), can be used for creating an **audio-based movie indexing scheme**. In addition, *clusters* of similar movies in terms of acoustic labelling could be populated.

References

1. S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Third Edition.* Academic Press, Inc., Orlando, FL, USA, 2008.

2. T. Giannakopoulos A. Pikrakis and S. Theodoridis. Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks. In *33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP08)*.
3. Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A multi-class audio classification method with respect to violent content in movies, using bayesian networks. In *IEEE International Workshop on Multimedia Signal Processing (MMSP07)*.
4. Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. Violence content classification using audio features. In *4th Hellenic Conference on Artificial Intelligence (SETN06)*.
5. Jeho Nam and Ahmed H. Tewfik. Event-driven video abstraction and visualization. *Multimedia Tools Appl.*, 16(1-2):55–77, 2002.
6. T. Giannakopoulos A. Pikrakis and S. Theodoridis. A computationally efficient speech/music discriminator for radio recordings. In *2006 International Conference on Music Information Retrieval and Related Activities (ISMIR06)*.
7. A. Pikrakis, T. Giannakopoulos, and S. Theodoridis. A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks. *Multimedia, IEEE Transactions on*, 10(5):846–857, 2008.
8. Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. Music tracking in audio streams from movies. In *IEEE International Workshop on Multimedia Signal Processing 2008 (MMSP08)*.
9. Chung-Hsien Wu. and Chia-Hsin Hsieh. Multiple change-point audio segmentation and classification using an mdl-based gaussian model. *IEEE Transactions on Audio, Speech and Language Processing*, 14:647–657, 2006.
10. M. Omar, U. Chaudhari, and G. Ramaswamy. Blind change detection for audio segmentation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2005.
11. T.N. Sainath, D. Kanevsky, and G. Iyengar. Unsupervised audio segmentation using extended baum-welch transformations. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2007.
12. A. Vasconcelos, N.; Lippman. Towards semantically meaningful feature spaces for the characterization of video content. In *International Conference on Image Processing, 1997*, pages 25–28.
13. N. V. Lobo A. Datta, M. Shah. Person-on-person violence detection in video data. In *IEEE International Conference on Pattern Recognition, Canada, 2002.*
14. Zeeshan Rasheed and Mubarak Shah. Movie genre classification by exploiting audio-visual features of previews. In *In Proceedings 16th International Conference on Pattern Recognition*, pages 1086–1089, 2002.
15. Y. Wang and L. Guan. Recognizing human emotional state from audiovisual signals. *Multimedia, IEEE Transactions on*, 10:936–946, 2008.
16. E.; Tsapatsoulis N.; Votsis G.; Kollias S.; Fellenz W. Cowie, R.; Douglas-Cowie and J. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18:32–80, 2001.
17. A. Hanjalic. Extracting moods from pictures and sounds: towards truly personalized tv. *Signal Processing Magazine, IEEE*, 23:90–100, 2006.