GUNSHOT DETECTION IN AUDIO STREAMS FROM MOVIES BY MEANS OF DYNAMIC PROGRAMMING AND BAYESIAN NETWORKS

Aggelos Pikrakis

Dept. of Informatics
University of Piraeus, Greece
e-mail: pikrakis@unipi.gr
URL: http://www.unipi.gr

Theodoros Giannakopoulos and Sergios Theodoridis

Dept. of Informatics and Telecommunications
University of Athens, Greece
e-mail: {tyiannak, stheodor}@di.uoa.gr
URL: http://www.di.uoa.gr/dsp

ABSTRACT

This paper treats gunshot detection in audio streams from movies as a maximization task, where the solution is obtained by means of dynamic programming. The proposed method seeks the sequence of segments and respective class labels, i.e., gunshots vs. all other audio types, that maximize the product of posterior class label probabilities, given the segments' data. The required posterior probabilities are estimated by combining soft classification decisions from a set of Bayesian Network combiners. Tests that have been performed on a large set of audio streams indicate that the proposed method yields high performance in terms of both precision and recall of detected gunshot events.

Index Terms— Gunshot Detection, BNs, Dynamic Programming

1. INTRODUCTION

The increasing availability of multimedia content over the last decade via numerous distribution channels has highlighted the need for efficient mechanisms to protect sensitive groups of the population imperative. To this end, efficient audio characterization techniques can assist the process of detecting violent scenes in audio-visual content, such as movies. Due to the fact that violence in movies is frequently correlated with gunshots events, we propose in this paper a method that detects gunshots in audio movie content.

Previous work in the more general field of violence detection has mainly focused on processing visual data, e.g., [1] and [2]. Audio content characterization has so far received less attention. To this end, in [3], the changes in the entropy of the energy envelop of the audio signal are exploited as a means to assist processing of visual data. In [4], eight audio features were investigated for discriminating between violent and non-violent sounds in the case of pre-segmented data, i.e., a manual segmentation stage was assumed prior to classification. The method in [5] presented a gunshot/scream detector for audio data recorded in public places, by taking a separate detection decision for each short-term frame. In [6], the authors of this paper presented a robust method for classifying pre-segmented audio data from movies, into six classes, i.e., speech, music, environmental sounds, gunshots, screams and fights. The last three of these classes were considered to represent cases of violent content, and as a result the method in [6] was also employed as a binary classification scheme for violent vs. non-violent audio content.

In order to circumvent the need for a manual segmentation stage, this paper formulates gunshot detection as a maximization task. In other words, the method seeks the sequence of segments and the respective class labels, i.e., gunshots vs. all other audio types, that maximizes the product of posterior (class label) probabilities, given the segments data. Since an exhaustive approach to this solution is unrealistic, we resort to dynamic programming to solve this maximization task. The potential of this type of formulation has been exploited by the authors in [7] in the context of speech/music discrimination of radio recordings.

In order to estimate the required posterior probabilities for this two class problem, i.e., gunshots vs all, we resort to the following scheme:

- Audio data are considered to belong to one of the following eight classes: Music, Speech, Gunshots, Fights, Screams and three classes of environmental sounds
- For each one of the above eight classes, a separate Bayesian Network (BN) combiner has been trained to yield binary classification decisions for the class vs all other classes problem. By its nature, each one BN returns the respective posterior class probability.
- The posterior probabilities returned by the eight BNs are then processed to yield an estimate of the posterior probabilities of the two class problem of gunshots vs all.

Previous work by the authors in [6] has set evidence that the combination of decisions taken from an *ensemble* of one-vs-all BNs outperforms a single gunshots-vs-all BN for the binary problem of gunshots vs all classification. Thus, this paper builds upon the experience obtained in [6] and [7] in order to formulate the gunshot detection problem as a maximization task and embed a reliable posterior probability estimator into the respective dynamic programming solution. Furthermore, in order to improve the accuracy of segment boundaries, a post-processing stage based on Bayesian Networks is employed in the end.

The paper is structured as follows: Section 2 describes the feature extraction stage; Section 3 formulates gunshot detection as a maximization task and provides a dynamic programming solution; The posterior probability estimator and related issues are given in Section 4; The datasets that we have used along with the method's performance are presented in Section 6. Finally conclusions are drawn in Section 7.

2. FEATURE EXTRACTION

At a first step, the audio stream is broken into a sequence of non-overlapping short-term frames (50 msecs long) and twelve features

are extracted per frame. The proposed feature set has been the result of extensive experimentation. A short description of the adopted features is given in Table 1. Briefly, we have used two time-domain quantities, namely the Energy Entropy and the Zero Crossing Rate. Furthermore, we have used two quantities based on the morphology of the signal spectrogram, the first three Mel Frequency Cepstral Coefficients, Spectral RollOff and the ratio of the Zero Pitch frames (pitch was estimated by means of autocorrelation). Finally, two quantities based on the chroma vector have also been used. For each measure quantity a different statistic, computed from respective segments, is used as a feature, as shown in Table 1. For a detailed description the reader is referred to [6]. For notational purposes, let $\mathbf{F} = \{O_1, O_2, \ldots, O_T\}$ be the resulting feature sequence, where T is the number of short-term frames and $O_t, t = 1 \ldots T$ stands for the t-th 12-dimensional feature vector.

	Feature	Statistic
o_1	Spectrogram-based feature	std
o_2	1st Chroma-based feature	mean
o_3	2nd Chroma-based feature	median
o_4	Energy Entropy	max
o_5	MFCC 2	std
o_6	MFCC 1	max
o_7	ZCR	mean
o_8	Sp. RollOff	median
o_9	Pitch	zero ratio
o_{10}	MFCC 1	max/mean
o_{11}	Spectrogram	max
o_{12}	MFCC 3	median

Table 1. Features and related statistics

3. GUNSHOT DETECTION AS A MAXIMIZATION TASK

In this stage, gunshot detection is treated as a maximization task, where the solution is obtained by means of dynamic programming. The basic idea is to define a function that returns a score given a sequence of segments and respective class labels. We choose as the segmentation sequence the one corresponding to the maximum score. To this end, for an arbitrary sequence of K segments, let

$$\{d_1, d_2, \ldots, d_{K-1}, d_K \equiv T\}$$

be the frame indexes that mark the end of each segment. Therefore, the k-th segment starts at frame index $d_{k-1} + 1$ and ends at frame index d_k (the first segment starts at the first frame and ends at frame index d_1). In addition, let c_k be the class label of the k-th segment (gunshots or "other") and $p(c_k \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\})$, be the posterior probability of class label c_k given the sequence of observations (feature sequence) of the k-th segment. We then form the following product function

$$J(K, \{d_1, d_2, \ldots, d_{K-1}, d_K\}, \{c_1, \ldots, c_K\}) \equiv$$
$$p(c_1 \mid \{O_1, \ldots, O_{d_1}\}) \prod_{k=2}^{K} p(c_k \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) \quad (1)$$

where independence between successive segments has been assumed. $J(.)$ is the product of posterior probabilities of the class labels given the within the segments data and needs to be *m*aximized over all possible values of K, $\{d_1, d_2, \ldots, d_K\}$ and $\{c_1, c_2, \ldots, c_K\}$. Since an exhaustive approach would amount to an excessive computational load, we resort to a dynamic programming solution. For a detailed description of the solution the reader is referred to [7], where we formulated speech/music discrimination of radio recordings as a maximization task as well. In this paper, we only provide the basic steps of the solution.

In order to proceed it is important to make the assumption that $T_{min} \leq d_k - d_{k-1} \leq T_{max}, k = 1 \ldots K$, i.e., the duration of segments is bounded by T_{min} and T_{max}. This assumption is not restrictive because long segments will simply be broken into a chain of shorter ones, which can be concatenated at a simple post-processing stage. The reason that this assumption is adopted is that $J(.)$ is a product of probabilities and as such, favors segmentation sequences consisting of a small number of segments, even though the respective posterior probabilities can be quite low.

As it is common with dynamic programming techniques [8], a grid is first constructed by placing the feature sequence on the x-axis and the two states, i.e., gunshots/"other" on the y-axis. This is shown in Figure 1, where *G* stands for gunshots and *O* stands for all the other types of audio data. Each node has a physical meaning, e.g.,

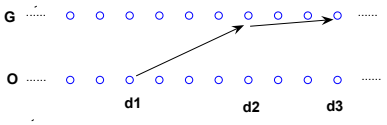

Fig. 1. Dynamic programming grid.

node (O_{d_k}, G), $T_{min} \leq d_k \leq T$ stands for the case that a segment which is part of a gunshot ends at frame index d_k. As a result, a path of K nodes corresponds to a possible sequence of segments and respective class labels. We have proved in [7] that the cost of a path of K nodes, say $\{(O_{d_1}, c_1), (O_{d_2}, c_2), \ldots, (O_{d_K}, c_K)\}$, is equal to the value of $J(.)$ for the respective segmentation sequence, provided that the transition cost, $T(.)$, between successive nodes is defined properly, i.e.,

$$T((O_{d_{k-1}}, c_{k-1}) \rightarrow (O_{d_k}, c_k)) = p(c_k \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) \quad (2)$$

Therefore, the best path on the grid maximizes $J(.)$ and corresponds to the desired segmentation solution. Details on computing the best-path sequence can be found in [7].

For the above formulation to have practical meaning, it is important that posterior probabilities are reliably estimated. As it will be presented in the next section, we have chosen to approximate $p(c_k \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\})$ by means of processing the decisions of an ensemble of Bayesian Network combiners.

4. POSTERIOR PROBABILITY ESTIMATION

As it was presented in Section 3, the proposed method requires the estimation of posterior probabilities for a two-class problem, i.e., gunshots vs all. Although we could have used one single BN for this task, our previous experience with pre-segmented data [6] suggests that a more complex estimator is needed. To this end, we have defined the following eight classes in order to describe audio content in movies: Gunshots, Fights, Screams, Music, Speech, "Others1", "Others2", "Others3". "Others1" consists of environmental sounds that have a relatively stable energy contour, e.g., wind, rain and low

energy noise. "Others2" refers to environmental sounds with abrupt changes in energy, e.g., moving objects, closing doors, etc. Finally, "Others3" covers sounds form various types of machinery, e.g., cars, airplanes, etc. The set of classes defined in this work is a refinement of the scheme proposed in [6]. The main deference is that the class "Others" in [6], has been decomposed into three classes, in order to achieve a more detailed description of movie content.

For each one of the eight classes, $\omega_l, l = 1, \ldots 8$, a binary classifier, i.e, a "One Versus All classifier " (OVAC) is employed. In this paper, each OVAC is a Bayesian Network combiner. Therefore, each OVAC estimates the posterior probability of the respective class, $p(\omega_l \mid \{O_{d_{k-1}+1}, F, O_{d_k}\})$, given the segment's data for each one of the eight binary problems. However, since the dynamic programming grid only requires the posterior probabilities of gunshots and "other" given the segment's data, the eight aforementioned probabilities are post-processed as follows:

- Let p_{max} be the maximum of the eight probabilities.
- If p_{max} was generated by the OVAC corresponding to gunshots, then in the grid $p(c_1 \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) = p_{max}$ and $p(c_2 \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) = 1 - p_{max}$
- In any other case, $p(c_1 \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) = 1 - p_{max}$ and $p(c_2 \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\}) = p_{max}$.

This type of post-processing leads to trusting the decision of the gunshots vs all OVAC only when it outperforms all other decisions, thus reducing the risk of error.

4.1. OVACs' architecture

Each OVAC is a Bayesian Network (BN) combiner and corresponds to a binary subproblem. All BNs follow the same architecture, shown in Figure 2. For the i-th subproblem, each input node, $R_{i,j}, j = 1, \ldots, 3$, corresponds to the binary decision of a simple k-NN classifier that operates on a 4-dimensional subspace. $R_{i,j} = 1$, if the input sample is classified to ω_i, and 0 otherwise. Y_i is the output node and corresponds to the true class label of the i-th binary subproblem. By inferring in the BN, we estimate the probability $p(\omega_i \mid \{O_{d_{k-1}+1}, \ldots, O_{d_k}\})$ with $P(Y_i = 1 | R_{i,1}, R_{i,2}, R_{i,3})$.

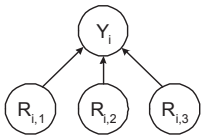

Fig. 2. BNC architecture for the i-th subproblem

We now describe how feature sequences of varying length are mapped to three 4-dimensional spaces on which the k-NN classifiers operate:

1. A statistic is computed for each one of the twelve feature sequences. The choice of statistics, e.g., average value, standard deviation, etc., was the result of extensive experimentation and further details can be found in [6]. This procedure leads to a 12-dimensional vector $\underline{v} = [v_1 \ldots v_{12}]$ for any audio segment.
2. Vector $\underline{v}$ is then broken into three 4-dimensional sub-vectors, each one of which is fed as input to a k-NN classifier.

In order to train the eight BNs more than 5000 audio segments have been used, which have been manually segmented and labelled from more than 30 movies. The length of segments varies in the range $[0.5 - 10]$ secs, with an average duration of 1.5 seconds. The fact that statistics are used also justifies the need for a minimum segment duration, i.e, T_{min}, as was stated in Section 3. In our study T_{min} and T_{max} were set equal to 0.5 and 1.5 secs respectively.

5. POST-PROCESSING FOR BOUNDARY CORRECTION

In order to further improve the system's accuracy, a post-processing scheme for boundary correction is applied on the segmented data. The idea behind this procedure is to maximize a probabilistic criterion related to the correctness of the boundary's position. This is performed with the following steps:

- Let T_b be the boundary (in secs) between two segments (gunshots and non-gunshots or vise versa). Furthermore, let c_{left} and c_{right} be the labels (1 for gunshots and 0 otherwise) of the segments on the left and the right of the boundary.
- Set $t = T_b - D$, where D is the *search range*, and $i = 0$.
- While $t \leq T_b + D$ do the following:
  - Let x_{left} be the signal in the range $[t - D, t]$.
  - Let x_{right} be the signal in the range $[t, t + D]$.
  - Using the gunshots vs all OVAC (say the g-th OVAC), estimate the probabilities $P_{left} = P(Y_g = c_{left} | x_{left})$ and $P_{right} = P(Y_g = c_{right} | x_{right})$. These are actually the probabilities that the left and right signals (x_{left} and x_{right}) around the current boundary position (t) satisfy the initial labels, i.e., c_{right} and c_{left}.
  - Set $P_i = P_{left} \cdot P_{right}$.
  - Set $i = i + 1$ and $t = t + 0.050$.
- Calculate $maxPos = \arg\max(P)$.
- Set the new boundary position as follows: $R = T + (maxPos \cdot 0.050 - D)$

This boundary correction algorithm improves system's performance if: a) the true boundary is indeed within the search range and b) the initial labels (c_{left} and c_{right}) are correct.

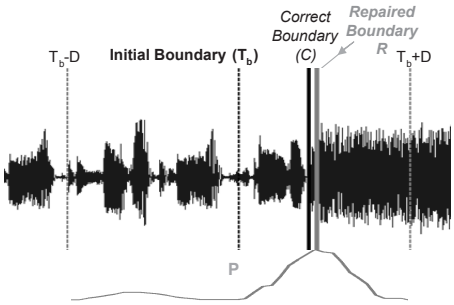

Fig. 3. Boundary Correction Algorithm: T_b is the center of the search region. R is found by maximizing P.

6. EXPERIMENTS - RESULTS

6.1. Datasets

Apart from the audio segments that were used for training the OVACs (see Section 4), a number of uninterrupted audio streams have been

recorded from more than 10 movies, in order to test the overall gunshot detection accuracy. For evaluation purposes, these audio streams were also manually segmented and labeled as either "Gunshot" or "Non-Gunshot". In total, almost two hours of audio data have been used.

6.2. Results

Two couples of measures that describe the system's performance have been calculated. The first refers to the total proportion of correctly classified data:

- Precision: The proportion of audio data that was classified as gunshots and was indeed gunshots.
- Recall: The proportion of gunshots data, that was correctly classified as gunshots.

The second couple of performance measures refers to the event detection ability of the algorithm:

- Detection Precision: The number of detected gunshots, that were indeed gunshots, divided by the total number of detected gunshots.
- Detection Recall: The number of correctly detected gunshots divided by the total number of true gunshots.

It has to be noted here, that by "correctly detected gunshots", we mean the detected gunshots that overlap with a true gunshot. The values of the two kinds of measures may differ a lot. In Figure 4, an example of gunshot detection is given (for an audio stream with two gunshots). In that case, the precision of classified data is $Pr = \frac{T/2+T}{T/2+1.2T} = \frac{1.5T}{1.7T} = 88.2\%$ and the recall is equal to $Re = \frac{T/2+T}{T+T} = \frac{1.5T}{2T} = 75\%$. The recall and precision measures of the detection performance of the method are both equal to 100%, since both of the existing gunshots were correctly detected, and no other gunshots were detected.

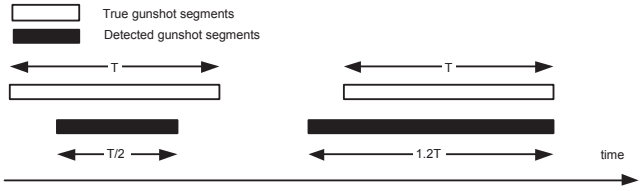

Fig. 4. Gunshot detection example

In Table 2 the values of the four performance measures are presented before and after the post-processing stage. As far as the detection performance of the algorithm is concerned, which is the most important in practice, the precision and recall have the same values before and after the post-processing stage, since only boundary correction is implied in this stage. As presented in the table, for the main stage of the proposed algorithm, the precision is 64% and the recall is 77.1%. Furthermore, according to the results, the post-processing technique improves the performance (0.5% for the precision rate and 0.7% for the recall rate).

As far as the event detection ability of the method is concerned, the precision rate is 78.8% and the recall is 90.6%. In other words, 9 out of 10 gunshots are detected, while for every 10 events that are detected almost 8 are indeed gunshots. The difference between the detection performance and the segmentation performance of the proposed algorithm is something expected. This actually implies that a large proportion of events are correctly detected, but an additional error is involved to the segmentation performance that pertains to the precise boundary specification.

	Stage 1	Stage2
Precision	64.0%	64.5%
Recall	77.1%	77.8%
Det. Precision	78.8%	
Rec. Precision	90.6%	

Table 2. System performance at different stages

7. CONCLUSIONS

An algorithm for gunshot detection in audio streams from movies has been presented. The problem has been treated as a maximization task, where the solution is obtained by means of dynamic programming, while the required posterior probabilities were estimated by combining classification decisions from a set of Baysian Network combiners. In terms of absolute duration, almost 65% of the detected data were indeed gunshots, while almost 80% of the gunshots data has been correctly detected. Furthermore, on an event-basis, the false alarm rate is almost 20%, while only 10% of the gunshots are not detected (false negative rate).

8. REFERENCES

[1] N. Vasconcelos and A. Lippman, "Towards semantically meaningful feature spaces for the characterization of video content," in *In International Conference on Image Processing, 1997*, vol. 1.

[2] N. V. Lobo A. Datta, M. Shah, "Person-on-person violence detection in video data," in *In IEEE International Conference on Pattern Recognition 2002*, Canada.

[3] Jeho Nam and Ahmed H. Tewfik, "Event-driven video abstraction and visualization," *Multimedia Tools Appl.*, vol. 16, no. 1-2, pp. 55–77, 2002.

[4] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis, "Violence content classification using audio features.," in *SETN*, 2006, pp. 502–507.

[5] Luigi Gerosa et al., "Scream and gunshot detection in noisy environments," in *15th European Signal Processing Conference (EUSIPCO-07), Sep. 3-7, Poznan, Poland*, 2007.

[6] 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*, 2007.

[7] Aggelos Pikrakis, Theodoros Giannakopoulos, and Sergios Theodoridis, "A dynamic programming approach to speech/music discrimination of radio recordings," in *15th European Signal Processing Conference (EUSIPCO-07), Sep. 3-7, Poznan, Poland*, 2007.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, 3d edition*, Academic Press, 2005.