Audio-based object recognition system for tangible acoustic interfaces

Fabio Antonacci, Giorgio Prandi,
Giuseppe Bernasconi, Roberto Galli,
Augusto Sarti
Politecnico di Milano
P.za Leonardo da Vinci, 32
I-20133 Milano, Italy
Email: antonacc/prandi/sarti@elet.polimi.it

***Abstract*—Tangible Acoustic Interfaces (TAIs) are innovative acoustic Human-Machine Interaction devices. Exploiting a number of contact sensors distributed on a surface, the vibrational signal generated from the interaction between the surface and an object moved by the user is acquired and analyzed to recognize what the user is doing on the device. The usage of vibrational sensors naturally opens the way also to classification and recognition applications. In this paper, a system to perform audio-based interaction object recognition is presented. The aim of the system is to recognize what object the human is using to interact with the TAI, by exploiting feature analysis and classification techniques. In particular, a frame-by-frame SVM-based classifier architecture is used to perform object recognition. The result is then filtered to eliminate the possible classification outliers. By training and testing our system using signals from four interaction objects at different Signal to Noise Ratios we have reached accuracies between 73% and 100% according to the object used, the quality of the acquired signal and the optional use of the classification filtering algorithm.**

I. Introduction

In the last years a big effort in the industry has been made on the development of intuitive and user-friendly Human Computer Interaction devices. If we categorize them based on the use of active or passive sensors, we can recognize Tangible Acoustic Interfaces (TAIs) in the latter class, since they use a set of accelerometers to acquire the vibrational signal produced by objects moved by the user on the TAI surface. Passive sensors present the intrinsic advantage of making it possible to transform whatever surface in a sensible one. In this context, solutions studied in TAI-CHI project [1] showed that an accurate and real-time localization of impulsive and continuous touches of different objects on passive surfaces (such as plexiglass and medium-density fiberboards) is possible. In particular, the solution proposed in [2] localizes the contact point analyzing of the Time Differences of Arrival (TDOAs) of the signal acquired by different sensors disposed on proper positions on the tangible surface. The only requirement of this system is that the interaction between the object and the board produces a noticeable sound.

In this paper we will enrich the framework used for the TAI-CHI project by adding the capability to recognize the object used for the interaction. This solution may be useful in several applications: one of the most interesting is the realization of interactive blackboards that behave in different ways according to the object used. In [3] we presented a solution for the object recognition which makes use of a fingerprinting approach. In particular, the Short Time Fourier Transform of the signal is processed to obtain a binary representation of the spectrum (the fingerprint). This signal is then used to discriminate different objects with a matching algorithm based on the minimization of the Manhattan distance. The promising results obtained in [3] encouraged us in developing a more refined system to improve the classification accuracy.

In this paper we present a solution that makes use of Support Vector Machines (SVMs) [4] to classify signals. SVMs cluster the feature space $\mathbb{R}^M$ in a number of regions divided by suitable separation surfaces that span a subspace of dimension $\mathbb{R}^{M-1}$. According to the region in which the feature vector under analysis fall, it is assigned to a specific class. The computation of the separation surfaces is done in a previous training stage. Support Vector Machines have been originally conceived as a binary classifier, but different extensions to classify multiple signals have been presented; a good overview may be found in [5]. In our work we make use of binary (two classes) SVMs, in a one-against-one configuration: in order to keep into account the case of N_O interacting objects, we make use of $N_O(N_O-1)/2$ binary SVM classifiers which compare the likelihood on object i vs. the likelihood of object j, $i \neq j$ and $i, j = 1, \ldots, N_O$. The global verdict is emitted on the base of the joint analysis of the verdicts of each classifier.

With a frame by frame analysis of the signal we compute a small set of features, based on waveform, spectrum and linear prediction coefficients of the signal, which are used for training and classification. The set of features has been selected starting from a larger set of 82 descriptors, through the exploitation of the Sequential Floating Forward Selection algorithm [6].

Since the path attenuation of the vibrational signals in thin solid surfaces is very high, in order to make the classification process more robust we select the signal acquired by the sensor closest to the contact point.

The rest of the paper is organized as follows: Section II presents the features used and the feature selection process. Section III discusses the proposed classification scheme. Section IV presents the experimental results obtained by testing

978-1-4244-4218-8/09/$25.00 ©2009 IEEE

the classification system in a real scenario. Finally Section V draws some conclusions.

II. Audio features

The recognition system uses a set of audio features extracted from the acquired audio signal to detect the object which moves on the tangible surface. Due to the huge amount of audio features described in literature, we used a feature selection approach over a large set of descriptors. The feature extraction and feature selection steps are described in the following sections.

A. Initial feature set

In our work, the initial large set of descriptors is composed of the 82 features listed in Table I.

Feature Type	Features	Ref.
Temporal	ZCR	[7]
Frequential	6 LPC coefficients.	[8]
Spectral	32 Audio Spectrum Envelope coefficients; Audio Spectrum Centroid; Audio Spectrum Spread; Bandwidth; 40 Quantized Spectral Coefficients.	[9][3]

TABLE I: Initial set of audio features.

In the following part of this section we will provide a short description for each feature. We will use the symbols $s(t), S(h), P(h)$ and $f(h)$ to denote, respectively, the time-domain signal, the amplitude and power spectrum of the signal and the frequency associated to the h-th bin of the signal spectrum (given the total number N_{ft} of spectral bins, we have $1 \leq h \leq N_{ft}$). For further details on specific features we invite the interested reader to consider the references in Table I.

1) Zero Crossing Rate: the Zero Crossing Rate (ZCR) counts the number of times the signal crosses the zero axes in each frame. In particular, the ZCR is computed as

$$ZCR = \frac{1}{2}\sum_{t=1}^{T-1} |\text{sign}[s(t)] - \text{sign}[s(t-1)]| \frac{F_s}{T} , \quad (1)$$

where F_s is the sampling frequency and T is the number of samples in each frame. If the signal is not zero-mean, a previous DC-offset removal stage should be implemented in order for ZCR to be informative.

2) Linear Prediction Coefficients: The linear prediction coefficients (LPC) refer to the theory of the Linear Prediction Coding that obtains an estimate $\hat{s}(t)$ of $s(t)$ as a linear combination of the past P samples:

$$\hat{s}(t) = \sum_{p=1}^{P} a_p s(t-p) . \quad (2)$$

The coefficients a_p are determined minimizing the Mean Square Error between $\hat{s}(t)$ and $s(t)$.

3) Audio Spectrum Envelope: The Audio Spectrum Envelope (ASE) is used in the literature to create a reduced version of the spectrogram. More specifically, after a logarithmic band subdivision, the value of ASE in a specific sub-band b is obtained by summing up the the values of $P(h)$ where $h \in b$.

4) Audio Spectrum Centroid and Audio Spectrum Spread: The Audio Spectrum Centroid (ASC) and Audio Spectrum Spread (ASS) make use of the logarithmic band subdivision defined in the previous paragraph. ASC and ASS compute the centroid and the spread of a modified version of the power spectrum of the signal as it were a probability density function:

$$ASC = \frac{\sum_{h'=0}^{N_{ft}/2-H_{low}} \log_2 \frac{f(h')}{1000} P(h')}{\sum_{h'=0}^{N_{ft}/2-H_{low}} P(h')} , \quad (3)$$

$$ASS = \sqrt{\frac{\sum_{h'=0}^{N_{ft}/2-H_{low}} (\log_2 \frac{f(h')}{1000} - ASC)^2 P(h')}{\sum_{h'=0}^{N_{ft}/2-H_{low}} P(h')}} , \quad (4)$$

where H_{low} is an user-defined value that limit the summation to the range of frequencies of interest and h' is the frequency bin index after the logarithmic band-subdivision.

5) Bandwidth: The bandwidth of the signal is obtained by the knowledge of ASC and $P(h')$ as follows:

$$B = \frac{\sum_{h'=0}^{N_{ft}/2} |ASC - f(h')| P(h')}{\sum_{h'=0}^{N_{ft}/2} P(h')} . \quad (5)$$

6) Quantized spectral coefficients: The computation of the quantized spectral coefficients is performed as follows: the average of the amplitude spectrum $S(h)$ is computed. Values of the signal $S(h)$ above the mean of $S(h)$ value are set to 1 in the quantized spectrum feature $S_{0/1}(h)$, while values below the mean value are set to 0.

B. Feature selection

The goal of the feature selection stage is to determine the optimal dimensionality of the vector (i.e. the value of l) and the related subset of features, with respect to some cost functions or classification results. Starting from the previous set of 82 features, it is possible to generate l-dimensional feature vectors, with $1 \leq l \leq 82$. It is desirable to keep the dimension l small to reduce the computational complexity and to avoid over-fitting problems in the training phase. In general, three methods have been proposed in the literature to perform feature selection [10][11]:

- *Filter method* - In the filter approach the feature selection algorithm is independent of any classifier; it filters out features that have a little chance to be useful in the classification task. The selection of the features is based on performance evaluation metrics computed directly from the data and does not take into account a direct feedback from classification results.
- *Wrapper method* - The wrapper method consists of evaluating a specific feature vector on the basis of classification results. The vector related to the best classification performance is chosen as the final result of the feature

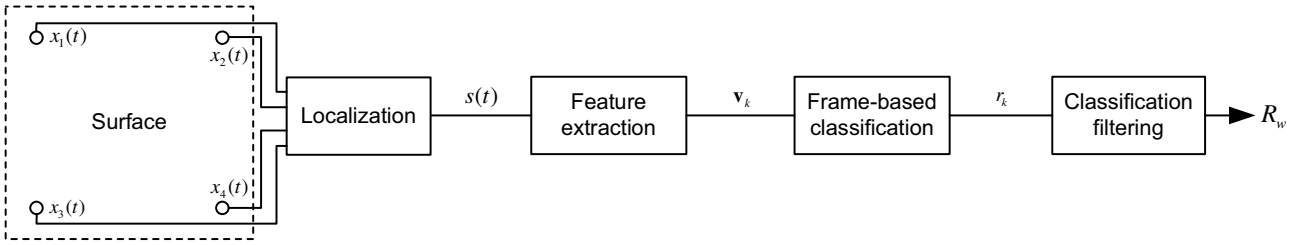

Fig. 1: Block diagram of the proposed interaction object recognition system

selection process. This approach tends to outperform filter methods, but at a much higher computational load.

- *Hybrid method* - The hybrid approach splits the problem of feature selection in two subproblems: the choice of the feature subset content, performed with a filter technique, and the selection of feature vector dimension, which is carried out in a wrapper fashion. This approach allows a considerable speedup in terms of resources needed for computation with respect to a pure wrapper approach, while giving good results for what concerns the overall classification performance.

In our framework the feature selection is performed only once. Computational complexity, as a consequence, does not concern us. For this reason we have resorted to the Sequential Floating Forward Selection method, a wrapper-based approach. Starting from an empty set of descriptors, the following two steps are iteratively performed in SFSS algorithm:

1) **First step**: a new feature l in the set L is added to the feature set currently used $\mathcal{Z}$. The feature added is chosen as the one that maximizes a predetermined cost function based on the accuracy obtained by the classifier:

$$l^+ = \arg \max_{l \in L - \mathcal{Z}} MCRR(\mathcal{Z} \cup l) \ , \qquad (6)$$

where $MCCR(\mathcal{Z} \cup l)$ is the Mean-Correct Classification Rate and it measures the average Correct Classification Rate of the cross-fold validation conducted using the set $\mathcal{Z} \cup l$;

2) **Second step**: among all the features in the set $\mathcal{Z}$ the feature l^- whose deletion best improves $MCCR(\mathcal{Z} - l^-)$ is discarded.

A user-defined number of iterations is performed (30 in our case) before stopping the algorithm.

III. Classification system

The general architecture of the recognition system is shown in Figure 1. The signal is acquired using an array of four microphones which are positioned on the corners of the surface. The four audio signals are used by the localization module to localize the contact point between the object and the surface; the signal $s(t)$ of the microphone closest to the contact point, localized using the algorithm in [2], is sent to the feature extraction module; in the feature extraction step, the signal is subdivided in small audio frames and the reduced set of features selected through the procedure described in Section II-B is extracted for each frame. This way, a feature vector $\mathbf{v}_k$ is associated to the k-th analyzed audio frame. Each feature vector $\mathbf{v}_k$ is then sent to the frame-based classification module which performs a frame-by-frame object recognition task exploiting a set of SVMs in one-against-one configuration. The result r_k of the recognition for the frame k is used in the filtering step that aggregates fixed-size windows of temporally consecutive results to give an improved recognition estimation R_w (where w is the current fixed-size window index). The details of frame-based classification and classification filtering steps are described in the following sections.

A. Frame-based classification

The classification system uses a set of one-against-one SVMs to recognize the object used to interact with the surface. A one-against-one SVM is a binary classifier: it tells what of the two considered classes a specific feature vector is belonging to. In our case, given a number N_O of objects to recognize $N_O(N_O-1)/2$ binary SVMs have to be implemented, one for each couple of classes. A generic feature vector $\mathbf{v}_k$ is classified by each SVM; then, a max-wins algorithm finds the final class of the current vector by selecting the most recurring class from the results given by the binary classifiers. The general scheme of the frame-based classification module is shown in Figure 2: the vector $\mathbf{v}_k$ is fed into each of the $N_O(N_O-1)/2$ binary classifiers, which work in parallel (in the block diagram, the generic class ω_i, with $1 \leq i \leq N_O$, is associated to the object i); the result of each classification is then used to detect the mode of the current classification task performed by SVMs; the class corresponding to the mode of the classification results is given as output r_k of the module.

B. Classification filtering

The results of the previous step may show some outliers due to noise or isolated classification errors: in order to drastically reduce the errors we adopt a window-based filtering technique: a window of S_W consecutive frame-based classification results is taken and the mode class of the window is detected. For each window w the final result R_w is given from the mode computed among classification results of the frames of the current window.

IV. Experiments

We have tested the proposed system by implementing it into a real scenario. This Section presents the experimental setup

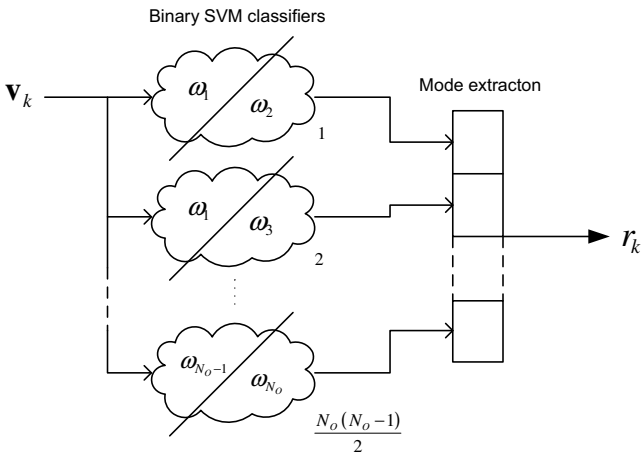

Fig. 2: Block diagram of the frame-based classification module

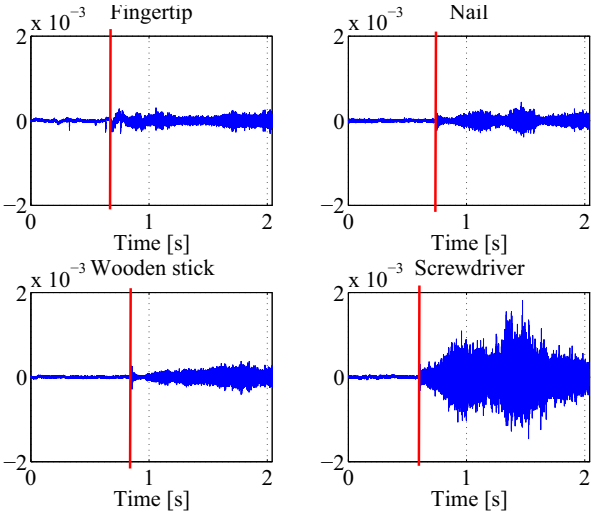

Fig. 3: Example of acquired signals for each interacting object. The vertical bold line separates noise from signal.

we have used to perform our tests and the experimental results obtained using four different interaction objects.

A. Experimental setup

We tested our system on a medium density fiber board. The rectangular sensible area has dimensions $60 \times 45 \times 0.8$ cm. At the corners of the sensible area, four Knowles BU 21771 accelerometers are mounted. A bi-adhesive tape ensures the contact between the sensors and the surface while preserving the signal transparency. The signal is acquired with a professional audio soundcard at a sampling rate of 16 kHz. The analysis frames for feature extraction are non-overlapped and 0.1 s long. Classification filtering analysis windows are non-overlapped as well. Later we will analyze the performance of the system as a function of the window size S_W.

To test our system, four interacting objects have been used, namely:

- Fingertip (F);
- Nail (N);
- Wooden stick (W);
- Screwdriver (S).

In some conditions, due to interferences in the power supply, we noticed the presence of a strong component in the signal at a frequency of 50 Hz and harmonics up to 500 Hz. A notch filter may optionally remove, when needed, this undesired component in the signal.

Figure 3 shows an example waveform for each object. At the beginning of each waveform a noise excerpt is present: the vertical bold line indicates the beginning of the useful signal. It can be observed that the fingertip produces a weak vibration that is likely dominated by noise. At the other extreme, the screwdriver produces a loud and noticeable vibration.

In order to perform feature selection, training and classification in different SNR conditions, we have considered the case of signal corrupted with 50 Hz noise and the case when the notch filter is present to remove the undesired components. Table II shows the SNR related to each object when the filter is activated (SNR_F) and the filter is not activated (SNR_{NF}). The object has been moved in proximity of one of the four sensors.

	Fingertip	Nail	Wooden stick	Screwdriver
SNR_F	4.93 dB	16.59 dB	22.49 dB	26.29 dB
SNR_{NF}	1.45 dB	10.50 dB	12.20 dB	16.20 dB

TABLE II: Average SNR values for each object.

B. Feature selection results

Using the configuration SNR_F, the original feature set has been reduced as depicted in Table III. Instead, using the configuration SNR_{NF}, the final set of features after feature selection is shown in Table IV.

Feature Type	Features
Temporal	ZCR.
Frequential	3rd and 5th LPC coefficients.
Spectral	Audio Spectrum Centroid; 1st and 8th Spectral coefficients; 2nd and 8th Audio Spectrum Envelope coefficients.

TABLE III: Remaining features after feature selection step, using filtered signals.

Feature Type	Features
Temporal	ZCR.
Frequential	2nd and 3rd LPC.
Spectral	Audio Spectrum Spread; 1st, 2nd, 11th and 21th Spectral coefficients; 2nd, 3rd, 8th and 18th Audio Spectrum Envelope coefficients.

TABLE IV: Remaining features after feature selection step, using non filtered signals.

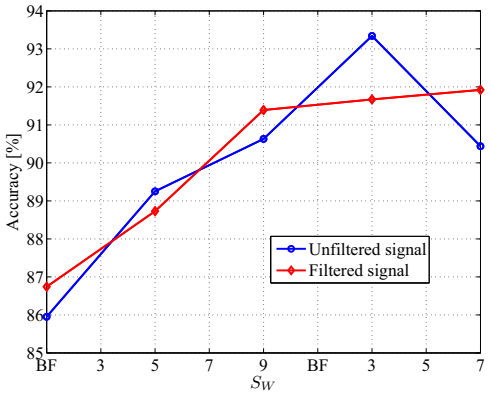

Fig. 4: Accuracy of the classification result as a function of the filtering window S_W. BF (Before classification Filtering) marks the $A_{r,\%}$ result obtained at the output of the frame-based classification module.

C. Classification results

Using the experimental setup described in previous section, we have conducted a set of experiments to evaluate the behavior of the classification system with filtered and unfiltered data, using an approach based on the method described in [3]. We have computed classification accuracy measure before and after the classification filtering to evaluate the performance of the frame-based classifier and the filtering block.

Given the total number N_f of analyzed signal frames, the accuracy before the classification filtering block is computed as follows:

$$A_{r,\%} = \frac{\text{correctly classified frames}}{N_f} \cdot 100 \qquad (7)$$

Given the size S_W, in number of frames, of the analysis window of the classification filtering block, the corresponding final accuracy is computed as follows:

$$A_{R,\%} = S_W \frac{\text{correctly classified windows}}{N_f} \cdot 100 \qquad (8)$$

Using these metrics, we have evaluated the average classification performance of our system. Results are shown in Figure 4. The classification filtering block increases the accuracy with respect to the results obtained from the frame-based classification module: in particular, it can be seen that for $S_W = 7$ a peak is detected for the accuracy of filtered data. Values of S_W higher than 9 (about one second of analysis) are not interesting for our real-time application due to the consequent increasing delay to obtain the classification result. Moreover, the performance of the classification filtering decreases in general for values of $S_W > 7$. It can be seen also that the accuracy for filtered and unfiltered data are very similar.

Following, the confusion matrices of the accuracy results before and after the classification filtering block ($S_W = 7$) are reported. In particular, Tables Va and Vb report the results obtained at the output of the frame-based classification block; Tables Vc and Vd, instead refer to the results obtained after applying the classification filtering. Null result stands for no decision: this happens when the mode of the classification results is not unique.

SNR_{NF}	F	N	W	S	Null
F	**99.00**	0.33	0.33	0	0.34
N	16.33	**80.00**	0	3.33	0.34
W	0.33	8.33	**76.00**	13.67	1.67
S	0	5.33	2.33	**92.00**	0.33

(a) Confusion matrix for A_r accuracy, with $N_f = 315$ and using the unfiltered signal and the related feature set.

SNR_F	F	N	W	S	Null
F	**96.19**	1.27	0.32	2.22	0
N	0	**73.33**	14.60	3.17	8.90
W	4.76	5.08	**80.00**	9.21	0.95
S	0	1.59	3.49	**94.29**	0.63

(b) Confusion matrix for $A_{r,\%}$ accuracy, with $N_f = 315$ and using the filtered signals and the related feature set.

SNR_{NF}	F	N	W	S	Null
F	**100.00**	0	0	0	0
N	15.56	**84.44**	0	0	0
W	0	2.22	**84.44**	13.34	0
S	0	2.22	0	**97.78**	0

(c) Confusion matrix for $A_{R,\%}$ accuracy, with $N_f = 315$, $S_W = 7$ and using the unfiltered signals and the related feature set.

SNR_F	F	N	W	S	Null
F	**100.00**	0	0	0	0
N	0	**84.44**	8.89	0	6.67
W	2.22	2.22	**91.12**	2.22	2.22
S	0	0	2.22	**97.78**	0

(d) Confusion matrix for $A_{R,\%}$ accuracy, with $N_f = 315$, $S_W = 7$ and using the filtered signals and the related feature set.

V. Conclusions

In this paper a system to perform audio-based object recognition system for acoustic tangible interfaces has been proposed. The core of the system is composed of a set of SVMs in one-against-one configuration. The global verdict is based with a Max-Wins policy among all the $N_O(N_O-1)/2$ verdicts. Finally, a classification filtering is used to attenuate the effect of isolated errors in the classification process. The features used have been selected with the Sequential Floating Forward Selection. Experimental results conducted with real data show that the system may reach an accuracy between 73% and 100%. The computational cost of the proposed algorithm is kept low by the use of Support Vector Machines. We are quite confident that the proposed algorithm may be implemented on a commercial DSP.

References

[1] "Tai-chi project," Mar. 2004, http://www.taichi.cf.ac.uk/.

[2] G.De Sanctis, D.Rovetta, A.Sarti, G.Scarparo, and S.Tubaro, "Localization of tactile interactions through TDOA analysis: Geometric vs. inversion-based method," in *Proceedings of 2006 European Signal Processing Conference (EUSIPCO 2006)*, Sept. 2006.

[3] F. Antonacci, L. Gerosa, A. Sarti, S. Tubaro, and G. Valenzise, "Sound-based classification of objects based on a robust fingerprinting approach," in *Proc. of 2007 European Signal Processing Conference, EUSIPCO 2007*, Sep. 2007.

[4] C.Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, 1995.

[5] J. Yang, X. Yang, and J. Zhang, "A parallel multi-class classification support vector machine based on sequential minimal optimization," in *IMSCCS '06: Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences - Volume 1 (IMSCCS'06)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 443–446.

[6] K.Z.Mao, "Fast orthogonal forward selection algorithm for feature subset selection," *IEEE Transactions on Neural Networks*, vol. 13, Sep 2002.

[7] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.

[8] P.Stoica and R.Moses, *An introduction to Spectral analysis*. Prentice Hall, 1997.

[9] H. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: audio content indexing and retrieval*. John Wiley & Sons, 2005.

[10] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[11] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *15th European Signal Processing Conference (EUSIPCO-07), Sep. 3-7, Poznan, Poland*, 2007.