Impact Sounds Classification for Interactive Applications via Discriminative Dictionary Learning

Christos Tzagkarakis*, Nikolaos Stefanakis*,† and George Tzagkarakis*
*Foundation for Research and Technology - Hellas, Institute of Computer Science, Heraklion, Greece
†Hellenic Mediterranean University, Dept. of Music Technology and Acoustics, Rethymno, Greece
{tzagarak, nstefana, gtzag}@ics.forth.gr

***Abstract*—Classification of impulsive events produced from the acoustic stimulation of everyday objects opens the door to exciting interactive applications, as for example, gestural control of sound synthesis. Such events may exhibit significant variability, which makes their recognition a very challenging task. Furthermore, the fact that interactive systems require an immediate response to achieve low latency in real-time scenarios, poses major constraints to be overcome. This paper focuses on the design of a novel method for identifying the sound-producing objects, as well as the location of impact of each event, under a low-latency assumption. To this end, a sparse representation coding framework is adopted based on learned discriminative dictionaries from short training and testing data. The performance of the proposed method is evaluated on a set of real impact sounds and compared against a nearest neighbour classifier. The experimental results demonstrate the high performance improvements of our proposed method, both in terms of classification accuracy and low latency.**

***Index Terms*—Impact sound classification, real-time processing, sparse representation classification, discriminative dictionary sparse coding**

I. Introduction

Impact sounds are a special category of non-stationary sounds, characterized by an abrupt onset and a rapid decay. Due to their high localization in time, they are excellent signals for conveying information, such as the time and location of an event's occurrence, or the size and material of the objects that collide. When produced by humans, these sounds can be naturally associated to gestures, which can be used to extract higher-level symbolic information, such as timing and rhythm, or to pass commands to a computational system for performing certain actions. For this reason, the accurate and robust detection and classification of impact sounds in an automatic fashion may serve a wide range of applications, from security and surveillance systems [1] to music information retrieval [2] and gesture recognition [3], [4].

In this paper, we examine the classification of impact sounds from the perspective of a gesture recognition system. Our goal is to process a monophonic acoustic signal, which is acquired by the natural stimulation of simple daily objects, in order to trigger a sound synthesis engine to produce a synthetic sound. This is a particularly challenging task considering the real-time constraints of this application. As it is stated in [5], the time between a gesture and its computer-generated audible reaction should be below ten milliseconds, which poses strict limitations to the amount of information that can be extracted from the sonic gesture before deriving a decision.

Even when impact sounds originate from the same object, slight variations in the impact force and location may result in a rich variability of their acoustic structure. In order to learn this variability, several training samples need to be acquired and processed to extract the most salient features that represent each class. To this end, principal component analysis (PCA) and independent component analysis (ICA) have been employed by [6] as redundancy reduction techniques, which seek for the directions that best describe the distribution of the data. In other works [7], [8], probabilistic models have been exploited for learning the spectral templates that are characteristic of each event. Finally, a much simpler approach was introduced in our previous work [9], demonstrating that instantaneous classification of such events is possible by employing very short segments of the acoustic signals in conjunction with a nearest neighbour (NN) classifier based on spectral features.

In this work, we investigate the challenging scenario of discriminating not only between different objects, but also between distinct impact locations for the same object. For illustration purposes, we consider the case of an empty bottle of beer, a plastic bucket and a box made of recycled paper. Motivated by our previous works on robust speaker identification using short training and testing data [10], [11], we jointly address the problem of object and impact location identification based on sparse representation coding.

More specifically, the problem of impact sounds classification is examined in light of a sparse representations framework, by relying on *discriminative dictionary sparse coding* techniques. The features that are given as input to our proposed classifiers are based solely on *spectral magnitude information* providing a more generic solution, since the additional incorporation of phase information could deteriorate the system's performance, in terms of recognition accuracy, due to its sensitivity in detecting the onset parameters. To the best of our knowledge, this is the first time that sparse representations are exploited for classifying impact sounds, whilst achieving high recognition rates in a real-time fashion, based only on spectral magnitude information.

This work has received funding from the European Union Horizon's 2020 research and innovation programme under grant agreement No. 780315 (SEMIoTICS) and the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement No. 2285 (neuronXnet).

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

We emphasize that, during the performance evaluation, we are interested in demonstrating the effectiveness of sparse representations based on learned dictionaries, and compare against the system described in our previous work [9], under a *limited amount of training* data. As a result, typical off-the-shelf classification methods requiring a great amount of training data are not examined during the comparative study. We also note that sparse representations have been previously used for sound/acoustic event detection and classification in [12], [13]. However, their fundamental difference with our proposed approach is that they exploit sparse representations as a feature extraction process, where the calculated sparse coefficients are given as input to a typical off-the-shelf classifier. On the contrary, this paper focuses on the design of a real-time impact sounds recognition system, which employs sparse representations as the classifier per se, whilst a fast feature learning process is adopted based solely on the spectral magnitude information.

The rest of the paper is organized as follows: Section II overviews the sparse representations framework, along with the NN classification method. The proposed impact sounds classification method is analyzed in Section III, followed by an experimental evaluation on real impact sound recordings. Finally, Section IV summarizes the main results and gives directions for further extensions.

II. Impact Sounds Classification Methods

As mentioned above, the main objective of this work is to extend our previous study [9] within the framework of real-time sparse representation classification and discriminative dictionary sparse coding of impact sounds using only the spectral magnitude information.

Following a similar feature extraction process as in [9], each impact sound is transformed in the frequency domain using a typical Fourier transform. Let us assume that during the training process a set of known acoustic events (an acoustic event is defined as the acoustic triggering of a specific object's region) is used to build a set of training data. The frequency domain representation $x_{i,j}(k)$, where $k = 0, \ldots, N-1$ denotes the frequency index and N the fast Fourier transform (FFT) size, corresponds to the j^{th} discrete time-domain signal of the i^{th} class. It is important to note that a small subset of consecutive frequency bins is considered, i.e., the frequency range of an impact sound need not be entirely spanned and it is rather sufficient to capture its most dominant acoustic modes by focusing on the frequency range $f_{min} \leq \frac{kFs}{N} \leq f_{max}$. The minimum f_{min} and maximum f_{max} frequency limit is the same for all classes, with F_s expressing the sampling rate of the time-domain signals. Under the aforementioned hypotheses, we may now denote the d-dimensional vector associated to an impact sound as

$$\mathbf{x}_{i,j} = [x_{i,j}(k_{min}), \ldots, x_{i,j}(k_{max})]^T \in \mathbb{R}^d, \qquad (1)$$

where $\mathbf{x}_{i,j}$ corresponds to the magnitude information of the computed FFT, while k_{min} and k_{max} is the smallest and largest index of the frequency bins that are taken into consideration, respectively.

Let us also assume that each d-dimensional vector is normalized to unit ℓ_2 norm and stored as the *spectral feature vector* representative of the j^{th} instance of the i^{th} class $\mathbf{v}_{i,j} := \mathbf{x}_{i,j}/\|\mathbf{x}_{i,j}\|_2$. All the normalized feature vectors $\{\mathbf{v}_{i,j}\}$ constitute the columns of a matrix, the so-called *dictionary*,

$$\mathbf{V}_i = [\mathbf{v}_{i,1}|\mathbf{v}_{i,2}|\ldots|\mathbf{v}_{i,n_i}] \in \mathbb{R}^{d\times n_i}. \qquad (2)$$

The total number of feature vectors corresponding to the training data, obtained from the impact sounds recordings, is equal to $N_{tr} = n_1 + \ldots + n_S$, where S is the total number of classes. The overall training data matrix $\mathbf{V}$, defined by

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_{1,1}|\cdots|\mathbf{v}_{1,n_1}|\mathbf{v}_{2,1}|\cdots|\mathbf{v}_{2,n_2}|\cdots|\mathbf{v}_{S,1}|\cdots|\mathbf{v}_{S,n_S}] \\ &= [\mathbf{V}_1|\mathbf{V}_2|\cdots|\mathbf{V}_S] \in \mathbb{R}^{d\times N_{tr}} \ , \qquad (3) \end{aligned}$$

corresponds to the concatenation of all the training data matrices $\mathbf{V}_i$, for $i = 1, \ldots, S$.

A. Classification of Impact Sounds via Nearest Neighbours

Let a NN classifier to be applied on the *test* feature vector $\mathbf{y}_t$, which is normalized to unit ℓ_2 norm. Then, $\mathbf{y}_t$ is compared against all the different class instances during the testing phase in order to find the class with the maximum correlation as follows,

$$\hat{i}_j = \arg\max_{i,j} |\langle \mathbf{v}_{i,j}, \mathbf{y}_t\rangle| \ , \ i = 1, \ldots, S \ , \ j = 1, \ldots, N_{tr} \ , \quad (4)$$

where $\langle \mathbf{a}, \mathbf{b}\rangle = \mathbf{a}^T\mathbf{b}$ denotes the inner product between two vectors and $\hat{i}_j$ carries the index of the selected class i (and optionally the instance index j).

B. Classification of Impact Sounds via Sparse Representations

This section overviews the method of sparse representation classification (SRC) [10] for impact sounds recognition. Let $\mathbf{y}_t$ be a normalized feature vector corresponding to the i^{th} class that can be expressed as a linear combination of the training samples associated with this class, as follows

$$\mathbf{y}_t = c_{i,1}\mathbf{v}_{i,1} + c_{i,2}\mathbf{v}_{i,2} + \cdots + c_{i,n_i}\mathbf{v}_{i,n_i} = \mathbf{V}_i\,\mathbf{c}_i \ , \qquad (5)$$

where the coefficients vector $\mathbf{c}_i = \{c_{i,j}\}_{j=1}^{n_i}$ represents the test vector $\mathbf{y}_t$ in terms of the columns of $\mathbf{V}_i$.

Furthermore, by combining (5) with (3), the test feature vector can be expressed in terms of the overall training data matrix $\mathbf{V}$, namely, $\mathbf{y}_t = \mathbf{V}\mathbf{c}$, where

$$\mathbf{c} = [0, \ldots, 0, c_{i,1}, c_{i,2}, \ldots, c_{i,n_i}, 0, \ldots, 0]^T \in \mathbb{R}^{N_{tr}} \qquad (6)$$

denotes the coefficients vector, hereafter called the *sparse code*. The elements of the sparse code $\mathbf{c}$ are all zero except for those associated with the training samples of the i^{th} class. An approximate sparse solution can be found by solving the following ℓ_1 norm optimization problem,

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \ \text{ s.t. } \ \mathbf{y}_t = \mathbf{V}\mathbf{c} \ . \qquad (7)$$

Given the test feature vector $\mathbf{y}_t$ and the training data matrix $\mathbf{V}$, the optimization problem (7) can be practically solved using several well-established algorithms. In this paper, we investigate the efficiency of the orthogonal matching pursuit (OMP) [14] and the least absolute shrinkage and selection operator (LASSO) [15] algorithms, in terms of their classification accuracy. The choice of these two methods is motivated by the good trade-off they achieve between a reduced computational complexity and an accurate sparse reconstruction.

The OMP algorithm, which belongs to the family of greedy sparse approximation methods, yields an approximation of the sparse code $\mathbf{c}$ by solving the following ℓ_2 norm optimization problem under an ℓ_0 norm constraint,

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{y}_t - \mathbf{V}\mathbf{c}\|_2^2 \ \text{ s.t. } \ \|\mathbf{c}\|_0 \leq K \ , \tag{8}$$

where K denotes the number of non-zero elements in $\hat{\mathbf{c}}$, which also determines the number of iterations of the algorithm.

LASSO is also adopted as a sparse approximation method for solving the optimization problem (7). In our case, there are many features that are highly correlated, especially with those coming from the same class, and LASSO tends to save one and ignore the rest of them. Furthermore, some features may not be reliable over time, for instance, due to the presence of noise. Elastic net regression [16] overcomes these limitations by combining the ℓ_1 and ℓ_2 norm penalties, as follows

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \frac{1}{2} \|\mathbf{y}_t - \mathbf{V}\mathbf{c}\|_2^2 + \lambda_1 \|\mathbf{c}\|_1 + \frac{\lambda_2}{2} \|\mathbf{c}\|_2^2 \ . \tag{9}$$

The objective of this method is to enhance stability and retain correlated features via the ℓ_2 norm penalty, whilst producing a sparse feature space through the ℓ_1 norm penalty.

Having estimated the sparse code $\mathbf{c}$, the classification is carried out by giving $\mathbf{c}$ as an input to a minimum reconstruction error classifier. Specifically, for a given class i, the test feature vector is estimated by $\hat{\mathbf{y}}_t = \mathbf{V}\delta_i(\hat{\mathbf{c}})$, and it is assigned to the class that gives the minimum residual between $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t = \mathbf{V}\delta_i(\hat{\mathbf{c}})$,

$$i^* = \arg\min_{i} \|\mathbf{y}_t - \mathbf{V}\delta_i(\hat{\mathbf{c}})\|_2^2 \ , \ i = 1, \ldots, S \ , \tag{10}$$

where $\delta_i : \mathbb{R}^{N_{tr}} \to \mathbb{R}^{N_{tr}}$ denotes an auxiliary indicator function for each class i. The entries of the vector $\delta_i(\hat{\mathbf{c}}) \in \mathbb{R}^{N_{tr}}$ are all zero except for those corresponding to the i^{th} class.

C. Classification of Impact Sounds via Discriminative Dictionary Sparse Coding

In this section, the method of discriminative dictionary sparse coding based on the Fisher discriminant criterion is introduced. This method was recently proposed in the framework of image recognition [17] and, to the best of our knowledge, it is applied for the first time to impact sounds classification.

Following the notation of Section II-B, the sparse representation optimization problem in (8) can be extended to the following dictionary sparse coding optimization problem,

$$\hat{\mathbf{D}}, \hat{\mathbf{C}} = \arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 \ \text{ s.t. } \ \|\mathbf{c}_j\|_0 \leq K \ , \tag{11}$$

for $j = 1, \ldots, N_{tr}$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\mathbf{D} \in \mathbb{R}^{d \times Z}$ is the learned dictionary, $\mathbf{C} \in \mathbb{R}^{Z \times N_{tr}}$ is the matrix of sparse codes, with $\mathbf{c}_j$ denoting the j^{th} column of $\mathbf{C}$, and Z corresponds to the dictionary size. We emphasize that the sparse codes $\{\mathbf{c}_j\}_{j=1}^{N_{tr}} \in \mathbb{R}^Z$ are of different dimensionality compared with the sparse code vectors introduced in Section II-B. However, the same symbol is used for notational convenience.

In order to enhance the discriminative capability of the estimated sparse codes, a dictionary sparse coding method based on the Fisher discriminative criterion [18] is introduced as follows

$$\hat{\mathbf{D}}, \hat{\mathbf{C}} = \arg\min_{\mathbf{D},\mathbf{C}} \mathcal{T}(\mathbf{V}, \mathbf{D}, \mathbf{C}) + \beta_1 \|\mathbf{C}\|_1 + \beta_2 \mathcal{R}(\mathbf{C}) \, ,$$
$$\text{s.t. } \|\mathbf{d}_j\|_2 = 1 \, , \ j = 1 \ldots, N_{tr} \ , \tag{12}$$

where $\mathcal{T}(\mathbf{V}, \mathbf{D}, \mathbf{C})$ is the discriminative data accuracy term, $\|\mathbf{C}\|_1 = \sum_{ij} |C_{ij}|$ is the sparsity penalty, $\mathcal{R}(\mathbf{C})$ is a discrimination term imposed on the sparse codes matrix $\mathbf{C}$, while β_1 and β_2 are regularization parameters controlling the trade-off between the discriminativeness of the training data, the sparsity level of the corresponding representation coefficients and the discriminative ability imposed on the sparse codes. Each column $\mathbf{d}_j$ of $\mathbf{D}$ is constrained to have a unit ℓ_2 norm. Notice also that the dictionary $\mathbf{D}$ should be capable of well representing each submatrix $\mathbf{V}_i$ (ref. (2), (3)), i.e., $\mathbf{V}_i = \mathbf{D}\mathbf{C}_i = [\mathbf{D}_1| \ldots |\mathbf{D}_i| \ldots |\mathbf{D}_S]\mathbf{C}_i = [\mathbf{D}_1| \ldots |\mathbf{D}_i| \ldots |\mathbf{D}_S][\mathbf{C}_i^1| \ldots |\mathbf{C}_i^i| \ldots |\mathbf{C}_i^S]^T = \mathbf{D}_1\mathbf{C}_i^1 + \ldots + \mathbf{D}_i\mathbf{C}_i^i + \ldots + \mathbf{D}_s\mathbf{C}_i^S$, where $\mathbf{V}_i \in \mathbb{R}^{d \times n_i}$, $\mathbf{D}_i \in \mathbb{R}^{d \times Z_i}$ and $\mathbf{C}_i^j \in \mathbb{R}^{Z_j \times n_i}$, with the size of $\mathbf{D}$ being $Z = Z_1 + \ldots + Z_j + \ldots + Z_S$. Furthermore, as long as the subdictionary $\mathbf{D}_i$ corresponds to the class i, we expect that $\mathbf{V}_i$ will be correctly represented by $\mathbf{D}_i$ but not by $\mathbf{D}_j$ for $i \neq j$. Therefore, the representation error $\left\|\mathbf{V}_i - \mathbf{D}_i\mathbf{C}_i^i\right\|_F^2$ will be small for the coefficients corresponding to $\mathbf{C}_i^i$, while $\|\mathbf{D}_j\mathbf{C}_i^j\|_F^2$ will be small for the coefficients $\mathbf{C}_i^j$. Given the above definitions, the discriminative accuracy term is defined by

$$\mathcal{T}(\mathbf{V}_i, \mathbf{D}, \mathbf{C}_i) =$$
$$\|\mathbf{V}_i - \mathbf{D}\mathbf{C}_i\|_F^2 + \left\|\mathbf{V}_i - \mathbf{D}_i\mathbf{C}_i^i\right\|_F^2 + \sum_{\substack{j=1 \\ j \neq i}}^{S} \left\|\mathbf{D}_j\mathbf{C}_i^j\right\|_F^2 \ . \tag{13}$$

The Fisher discriminant criterion is also introduced towards increasing the discriminative power of the sparse codes $\mathbf{C}$. More specifically, we aim at minimizing the within-class scatter $S_w(\mathbf{C}) = \sum_{i=1}^{S} \sum_{\mathbf{c}_j \in \mathbf{C}_i} (\mathbf{c}_j - \boldsymbol{\mu}_i)(\mathbf{c}_j - \boldsymbol{\mu}_i)^T$ between the sparse codes $\mathbf{C}$, whilst maximizing the between-class scatter $S_b(\mathbf{C}) = \sum_{i=1}^{S} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$, where $\boldsymbol{\mu}_i$ denotes the mean vector of $\mathbf{C}_i$, $\boldsymbol{\mu}$ is the mean vector of $\mathbf{C}$ and n_i is the number of training samples which belong to the i^{th} class.

After some algebraic manipulation (ref. [19]) the Fisher discriminant criterion is expressed as

$$\mathcal{R}(\mathbf{C}) = \text{tr}(S_w(\mathbf{C})) - \text{tr}(S_b(\mathbf{C})) + \tau \|\mathbf{C}\|_F^2 \, , \tag{14}$$

where τ is a parameter associated with the stability term $\|\mathbf{C}\|_F^2$. By inserting (13) and (14) into (12) we have that the discriminative sparse coding optimization problem based on the Fisher criterion is given by

$$\hat{\mathbf{D}}, \hat{\mathbf{C}} = \arg\min_{\mathbf{D},\mathbf{C}} \sum_{i=1}^{S} \|\mathbf{V}_i - \mathbf{D}\mathbf{C}_i\|_F^2 + \left\|\mathbf{V}_i - \mathbf{D}_i\mathbf{C}_i^i\right\|_F^2$$
$$+ \sum_{\substack{j=1 \\ j \neq i}}^{S} \left\|\mathbf{D}_j\mathbf{C}_j^i\right\|_F^2 + \beta_1 \|\mathbf{C}\|_1$$
$$+ \beta_2 \Big(\text{tr}(S_w(\mathbf{C})) - \text{tr}(S_b(\mathbf{C})) + \tau \|\mathbf{C}\|_F^2 \Big)$$
$$\text{s.t. } \|\mathbf{d}_j\|_2 = 1 \, , \ j = 1 \ldots, N_{tr} \ . \tag{15}$$

An approximate solution to the optimization problem (15) can be found by optimizing the dictionary $\mathbf{D}$ and the sparse codes $\mathbf{C}$ alternatively, i.e., updating $\mathbf{D}$ with $\mathbf{C}$ fixed, and then updating $\mathbf{C}$ by fixing $\mathbf{D}$.

Having obtained a solution of (15), the estimated dictionary $\hat{\mathbf{D}}$ and sparse codes $\hat{\mathbf{C}}$ are employed to perform the final classification. Given a normalized test feature vector $\mathbf{y}_t$, first we compute its sparse representation by solving

$$\hat{\gamma} = \arg\min_{\gamma} \left\| \mathbf{y}_t - \hat{\mathbf{D}}\boldsymbol{\gamma} \right\|_2^2 + \alpha \left\| \gamma \right\|_1 , \qquad (16)$$

where α is a regularization parameter controlling the sparsity level. The classified impact region associated with the query signal $\mathbf{y}_t$ is given by the minimum error metric as follows,

$$i^* = \arg\min_{i} \left\| \mathbf{y}_t - \mathbf{D}_i \hat{\gamma}_i \right\| + \theta \left\| \hat{\gamma} - \boldsymbol{\mu}_i \right\|_2^2 , \ i = 1, \ldots, S. \ (17)$$

This task is accomplished by finding the index that corresponds to the smallest error given by (17), where the first term is the reconstruction error of the class i, the second term is the distance between the coefficients vector $\hat{\gamma}$ and the learned mean vector $\boldsymbol{\mu}_i$ of the class i, and θ is a parameter used to balance the contribution of the two terms.

III. Experimental Evaluation

In this section, the classification accuracy of the methods proposed herein, namely, the sparse representation approach described in Section II-B and the discriminative sparse coding method discussed in Section II-C, are compared against the NN-based counterpart introduced in [9]. The experimental evaluation is carried out on three distinct objects: (i) an empty bottle of beer, (ii) a plastic bucket (originally used as a garbage bin) and (iii) an old cassette-case made of recycled paper. Hereafter, we will refer to these objects as the bottle, the bucket and the box, respectively. The bottle is excited with the help of a thin metallic rod in three regions, whereas both the bucket and the box are excited with the fingers of both hands of the user in four regions, respectively. The impact region of each object defines a class, thus the total number of classes is eleven. All the recordings took place in a room of dimensions $8 \times 7 \times 2.5$ m using a cardioid dynamic microphone (Shure SM 58) plugged into an external USB sound card. The dataset is publicly available at https://zenodo.org/record/2563718#.XHqIcIgzbIU.

The sampling frequency was set at 44100 Hz but the audio data was downsampled at 22050 Hz before the spectral processing. The simulations were conducted on a 64-bit MATLAB R2015a programming environment using a quad-core Intel Core i7-2670 with 8GB RAM. The training and testing data were automatically extracted from the corresponding audio files by using the onset detection algorithm analyzed in [9] (see Section 3.1 therein). More specifically, long audio files were segmented into multiple smaller files containing a single impact sound each. About 35 to 50 instances were recorded for each object and impact region, where for the bucket and the box, strikes from both fingers and hands were recorded, when applicable. Besides, for all impact regions, we have tried to produce different intensity levels in order to cover a wide dynamic range. Due to space limitations, more information on the excitation processes and recordings can be found in [9].

TABLE I
Mean correct classification rates of NN vs. SRC-OMP vs. SRC-LASSO vs. DDSC averaged over all classes. The length of the processing window is 20 msec.

# train. signals per class	**NN**	**SRC OMP**	**SRC LASSO**	**DDSC**
all	94.66	95.76	95.69	**95.94**
20	93.93	**95.02**	94.98	95.00
15	94.37	95.53	95.26	**96.42**
10	93.07	93.88	93.30	**94.47**
5	91.21	92.24	92.13	**92.30**
2	86.98	87.92	**88.54**	88.04

During the spectral analysis step (ref. Section II), the length of the processing window was set to 20, 9, 6 and 4 msec, respectively. The size of the FFT equals the length of the processing window. The impulsive nature of impact sound signals led us use the most informative frequency content ranging from 0 to 5000 Hz. Following a cross-validation procedure for all the models' parameters, the sparsity threshold K was chosen to be 5 during the SRC-OMP evaluation procedure, while for the SRC-LASSO classifier the regularization parameters λ_1 and λ_2 were set to 10^{-3} and 10^{-5}, respectively. For the discriminative dictionary sparse coding (DDSC) method, the parameters of the optimization problem (15) were set to $\beta_1 = 0.05$ and $\beta_2 = 10^{-3}$, respectively, while during the classification step (16)-(17) we set $\alpha = 0.01$ and $\theta = 10^{-3}$.

TABLE II
SRC-OMP vs. SRC-LASSO computation time in msec.

proces. win. (msec)	**SRC-OMP**	**SRC-LASSO**
20	0.3699	0.5163
9	0.3267	0.4543
6	0.3092	0.4273
4	0.3021	0.4185

The average correct classification rates are computed as the percentage of the correctly identified test impact sounds over the total number of test impact sounds per class. Fig. 1 depicts eleven groups of four bars, where the first bar corresponds to the NN, the second and the third bar indicate the SRC-OMP and SRC-LASSO, respectively, while the fourth bar corresponds to the DDSC method. Clearly, both the SRC and DDSC outperform the NN classifier for the majority of the impact regions.

This demonstrates the efficiency of DDSC and SRC to distinguish correctly not only the object's type but also the region being hit. By averaging all the mean correct classification rates, shown in Table I, reveals that DDSC achieves a higher accuracy than SRC if we use a training data matrix that consists of all the signals per class, in addition to 20, 15, 10, 5 and 2 signals per class, respectively. In particular, DDSC accomplishes approximately at least 94.5%, on average, for a training data matrix with more than 10 signals per class. Additionally, we observe that the highest accuracy of DDSC when using 2, 5, 10, 15 and 20 training signals per class is 96.42%, while NN and SRC achieve an inferior performance of 93.93% and 95.02%, respectively.

As mentioned already, another issue which is worth of in-

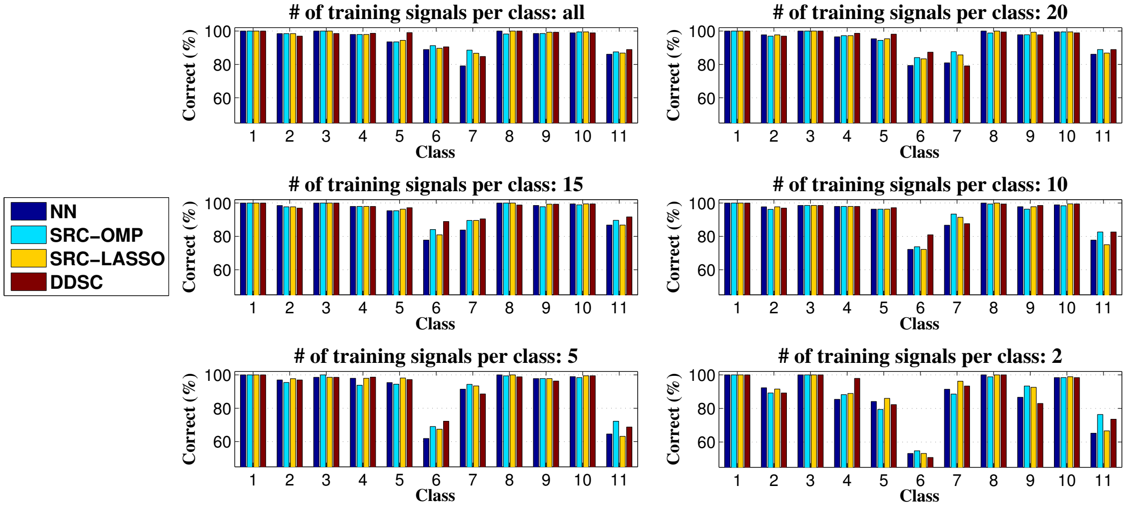

Fig. 1. Correct impact sounds classification rates as a function of the class index using an 20 msec processing window. The amount of training data equals (i) all the available training signals, (ii) 20 signals, (iii) 15 signals, (iv) 10 signals, (v) 5 signals, and (vi) 2 signals, respectively.

vestigating is the response speed of the classification system in a potential object hit. A classification method is characterized as real-time if the time for making a decision is less than the processing window length (in msec). Table II shows the computation times in msec for SRC-OMP and SRC-LASSO and for all window lengths. Clearly, both approaches achieve low latencies for all the distinct tasks and can be characterized as "real-time". Furthermore, we deduce that SRC is very useful when low latency is of great importance, whereas DDSC is more suitable for offline applications of impact sound recognition, such as labelling. As a final conclusion, we could state that SRC-LASSO ensures both a fast response and high classification rates. On the other hand, DDSC can be applied for labelling of impact sounds in an offline mode, whenever it is more significant to automatically label the recorded impact sounds, and thus avoid manual labelling operations.

IV. Conclusions

The sparse representation classification (SRC) and discriminative dictionary sparse coding (DDSC) methods were proposed for the classification of impact sounds using a limited amount of training data. An extensive performance evaluation on real impact sound recordings revealed their high performance in terms of classification accuracy. Furthermore, SRC better suits real-time applications, while DDSC is more appropriate for offline labelling tasks. As a future work, it is of great interest to examine the proposed methods with a larger set of daily objects.

References

[1] K. Łopatka, P. Zwan, and A. Czyżewski, "Dangerous sound event recognition using support vector machine classifiers," *Advances in Intelligent and Soft Computing*, vol. 80, pp. 49–57, 2010.

[2] K. Yoshii, M. Goto, and H. G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," in *Proceedings of 5th Int. Conf. on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 184–191.

[3] B. Zamborlin, F. Bevilacqua, M. Gillies, and M. D'Inverno, "Fluid gesture interaction design: applications of continuous recognition for the design of modern gestural interfaces," *ACM Transactions on Interactive Intelligent Systems*, vol. 3(5), 2014, article 22.

[4] M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92(4), pp. 632–644, 2004.

[5] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers," *Computer Music*, vol. 26(3), pp. 11–22, 2002.

[6] S. Cavaco and M. S. Lewicki, "Statistical modeling of intrinsic structures in impact sounds," *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3558–3568, 2007.

[7] U. Şimşekli, A. Jylhä, C. Erkut, and T. Cemgil, "Real-time recognition of percussive sounds by a model-based method," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–14, 2011.

[8] E. Battenberg, V. Huang, and D. Wessel, "Toward live drum separation using probabilistic spectral clustering based on the Itakura-Saito divergence," in *AES 45th Int. Conf.*, Helsinki, Finland, 2012.

[9] N. Stefanakis, Y. Mastorakis, and A. Mouchtaris, "Instantaneous detection and classification of impact sound: turning simple objects into powerful musical control interfaces," in *11th Sound and Music Computing Conference joint with 40th International Computer Music Conference*, Athens, Greece, September, 2014.

[10] C. Tzagkarakis and A. Mouchtaris, "Robust text-independent speaker identification using short test and training sessions," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, August 2010.

[11] C. Tzagkarakis and A. Mouchtaris, "Sparsity based robust speaker identification using a discriminative dictionary learning approach," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, August 2013.

[12] M. Zhang, W. Li, L. Wang, J. Wei, Z. Wu, and Q. Liao, "Sparse coding for sound event classification," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, October 2013, pp. 1–5.

[13] X. Lu, P. Shen, Y. Tsao, C. Hori, and H. Kawai, "Sparse representation with temporal max-smoothing for acoustic event detection," in *INTERSPEECH*, Dresden, Germany, September 2015, pp. 1176–1180.

[14] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Information Theory*, vol. 53(12), pp. 4655–4666, December 2007.

[15] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[16] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[17] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. Journal of Computer Vision*, vol. 109, pp. 209–232, September 2014.

[18] R. Duda, P. Hart, and D. Stork, *Pattern classification*, New York: Wiley-Interscience (2nd ed.), 2000.

[19] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, USA, June 2007, pp. 1–8.