

Wavelet-based acoustic detection of moving vehicles

Amir Averbuch · Valery A. Zheludev · Neta Rabin ·
Alon Schclar

Received: 28 March 2007 / Revised: 18 April 2008 / Accepted: 24 April 2008
© Springer Science+Business Media LLC 2008

Abstract We propose a robust algorithm to detect the arrival of a vehicle of arbitrary type when other noises are present. It is done via analysis of its acoustic signature against an existing database of recorded and processed acoustic signals to detect the arrival of a vehicle of arbitrary type when other noises are present. To achieve it with minimum number of false alarms, we combine a construction of a training database of acoustic signatures signals emitted by vehicles using the distribution of the energies among blocks of wavelet packet coefficients with a procedure of random search for a near-optimal footprint. The number of false alarms in the detection is minimized even under severe conditions such as: the signals emitted by vehicles of different types differ from each other, whereas the set of non-vehicle recordings (the training database) contains signals emitted by planes, helicopters, wind, speech, steps, etc. The proposed algorithm is robust even when the tested conditions are completely different from the conditions where the training signals were recorded. The proposed technique has many algorithmic variations. For example, it can be used to distinguish among different types of vehicles. The proposed algorithm is a generic solution for process control that is based on a learning phase (training) followed by an automatic real-time detection.

Keywords Wavelet · Wavelet packet · Acoustics signature · Classification · CART

1 Introduction

The goal is to detect the arrival of vehicles of arbitrary types such as various cars, vans, jeeps and trucks via the analysis of their acoustic signatures with minimal number of false alarms. This processing is done against an existing database of recorded acoustics signals. The problem is complex because of the great variability in vehicles types and because of the surrounding conditions that may contain sounds emitted by planes, helicopters, speech, wind, steps, to name a few, in the recorded database (training datasets). In addition, the velocities

A. Averbuch (✉) · V. A. Zheludev · N. Rabin · A. Schclar
School of Computer Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
e-mail: amir@math.tau.ac.il

of the vehicles, distances from the receiver, the roads the vehicles traveled on are highly variable as well and thus affect the recorded acoustics.

A successful detection depends on the constructed acoustics signatures that were built from characteristic features. These signatures enable to discriminate between vehicle (V) and non-vehicle (N) classes. Acoustics signals emitted by vehicles have quasi-periodic structure. It stems from the fact that each part of the vehicle emits a distinct acoustic signal which contains in the frequency domain only a few dominating bands. As the vehicle moves, the conditions are changed and the configuration of these bands may vary, but the general disposition remains. Therefore, we assume that the acoustic signature for the class of signals emitted by a certain vehicle is obtained as a combination of the inherent energies in the blocks of the wavelet packet coefficients of the signals, each of which is related to a certain frequency band. This assumption has been corroborated in the detection and identification of a certain type of vehicles (Averbuch et al. 2001a,b). The experiments in the current paper demonstrate that a choice of distinctive characteristic features that discriminate between V and N classes can be derived from blocks of wavelet packet coefficients. Extraction of characteristic features (parameters) is a critical task in the training phase of the process.

In order to identify the acoustic signatures, in the final phase of the process we combine the outputs from two classifiers. One is the well known Classification and Regression Trees (CART) classifier (Breiman et al. 1993). The other classifier is based on the comparison of distances between the test signal and sets of pattern signals from the V and N classes.

The paper has the following structure: In Sect. 2 we briefly review some works related to the subject of the present paper. The structure of the available data is described in Sect. 3. In Sect. 4 we outline the scheme of the algorithm and in Sect. 5 we describe in full details. Section 6 is devoted to presentation of the experimental results. Section 7 provides some discussion. Appendix 1 outlines the notions of the wavelet and wavelet packets transforms and Appendix 2 provides a detailed description of the random search for a near-optimal footprint (RSNOFP) method.

2 Related work

Several papers have been published in the area of separation between V and N sounds. Most of them describe systems for a military context.

Choe et al. (1996) extracted the acoustic features by using discrete wavelet transform. The feature vectors were compared with the reference vectors in the database using statistical pattern matching to determine the type of vehicle from where the signal originated. In Eom (1999), discrete cosine transform was applied to the signals and a time-varying autoregressive modeling approach was used for their analysis. Averbuch et al. (2001a) designed a system that is based on wavelet packets coefficients in order to discriminate between different types of vehicles. CARTs were used for classification of new unknown signals. In a later paper Averbuch et al. (2001b) used similar methods with multiscale local cosine transform applied on the frequency domain of the acoustic signal. The classifier that was used was based on *Parallel Coordinates* methodology. Wu et al. (1999) used the *eigenfaces method* (Sirovich and Kirby 1987), which was originally used for human face recognition, to distinguish between different vehicle sound signatures. The data was sliced into frames—short series of time slices. Each frame was then transformed into the frequency domain. Classification was done by projecting new frames on the principal components that were calculated for a known training set. Munich (2004) compared between several speech recognition techniques for classification of vehicle types. These methods were applied on short term Fourier transform of the vehicles' acoustic signatures.

3 The structure of the acoustics signals

The recordings were taken under very different conditions in different dates. The recordings sampling rate (SR) was 48,000 samples per second (SPS). It was downsampled to SR of 1,000 and 600 SPS.

We extracted from the set of recordings, which were used for training the algorithm, fragments that contain sounds emitted by vehicles and stored them as the V-class signals. Recorded fragments that did not contain vehicles sounds were stored as the N-class signals. Both classes were highly variable. Recordings in the V-class were taken from different types of vehicles during different field experiments under various surrounding conditions. In particular, the velocities of the vehicles and their distances from the recording device were varied. Moreover, the vehicles traveled on either various paved (asphalt) or unpaved roads, or on a mixture of paved and unpaved roads. Recordings in N-class comprised of sounds emitted by planes, helicopters, sometimes strong wind and speech nearby the receiver, to name a few.

Figure 1 displays portions of acoustic signals emitted by two cars with their Fourier transforms.

Figure 2 displays portions of acoustic signals emitted by a truck and a van with their Fourier transforms.

We see that the spectra of different cars differ from each other. It is even more apparent in the spectra of other vehicles. Figure 3 displays portions of acoustic signals emitted by a plane and a helicopter with their Fourier transforms, whereas Fig. 4 does the same for speech and wind patterns.

We realized that even within the same class (V or N), the signals differ significantly from each other. The same is true for their Fourier transforms. However, there are some common properties to all these acoustic signals that were recorded from moving vehicles. First, these signals are quasi-periodic in the sense that there exist some dominating frequencies in each signal. These frequencies may vary as motion conditions are changed. However, for the same vehicle, these variations are confined in narrow frequency bands. Moreover, the relative locations of the frequency bands are stable (invariant) to some extent for signals that belong to the same vehicle.

Therefore, we conjectured that the distribution of the energy (or some energy-like parameters) of acoustics signals that belong to some class over different areas in the frequency domain, may provide a reliable characteristic signature for this class.

4 Formulation of the approach

Wavelet packet analysis (see Appendix 1) is a highly relevant tool for adaptive search for valuable frequency bands in a signal or a class of signals. Once implemented, the wavelet packet transform of a signal yields a huge (redundant) variety of different partitions of the frequency domain. The transform is computational efficient. Due to the lack of time invariance in the multiscale wavelet packet decomposition, we use the whole blocks of wavelet packet coefficients rather than individual coefficients and waveforms. The collection of energies in blocks of wavelet packet coefficients can be regarded as an averaged version of the Fourier spectrum of the signal, which provides more sparse and more robust representation of signals compared to the Fourier spectrum. We can see it, for example, in Fig. 5. This figure displays the energies in the blocks of wavelet packet coefficients of the sixth level of a car acoustics signal. Here, the wavelet packet transform based on the orthogonal splines of sixth order is applied.

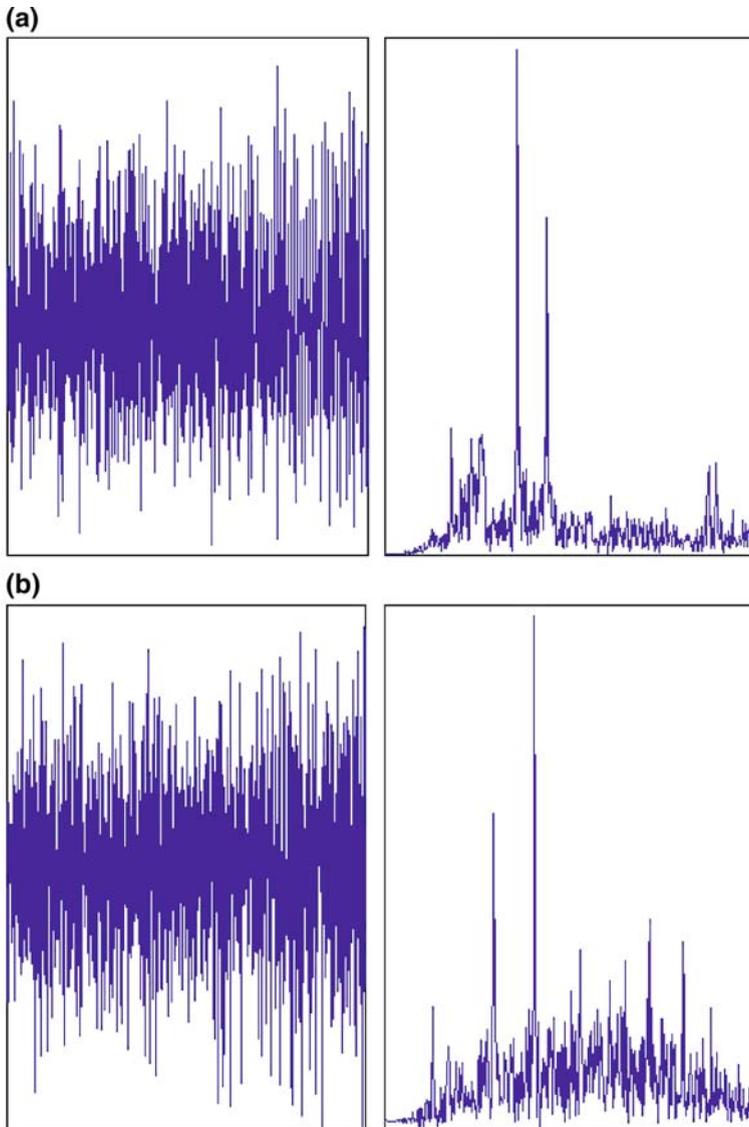


Fig. 1 Fragments of two car recordings and their spectra. Frames from left to right: first car, its spectrum; second car, its spectrum

Variation on the Best Basis algorithm (Coifman et al. 1992; Wickerhauser 1994) that searches for a few blocks that mostly discriminate a certain vehicle from other vehicles and the background was used in Averbuch et al. (2001a,b). Here, this approach did not prove to be robust and efficient because of the variability in vehicles types in Class V and kinds of the background in Class N recordings. Therefore, another way to utilize the wavelet packet coefficients blocks was chosen. This method can be characterized as a random search for the RSNOFP of a class of signals. This is close to the compressed sensing (Donoho 2006; Donoho and Tsai 2006; Candes et al. 2006) idea.

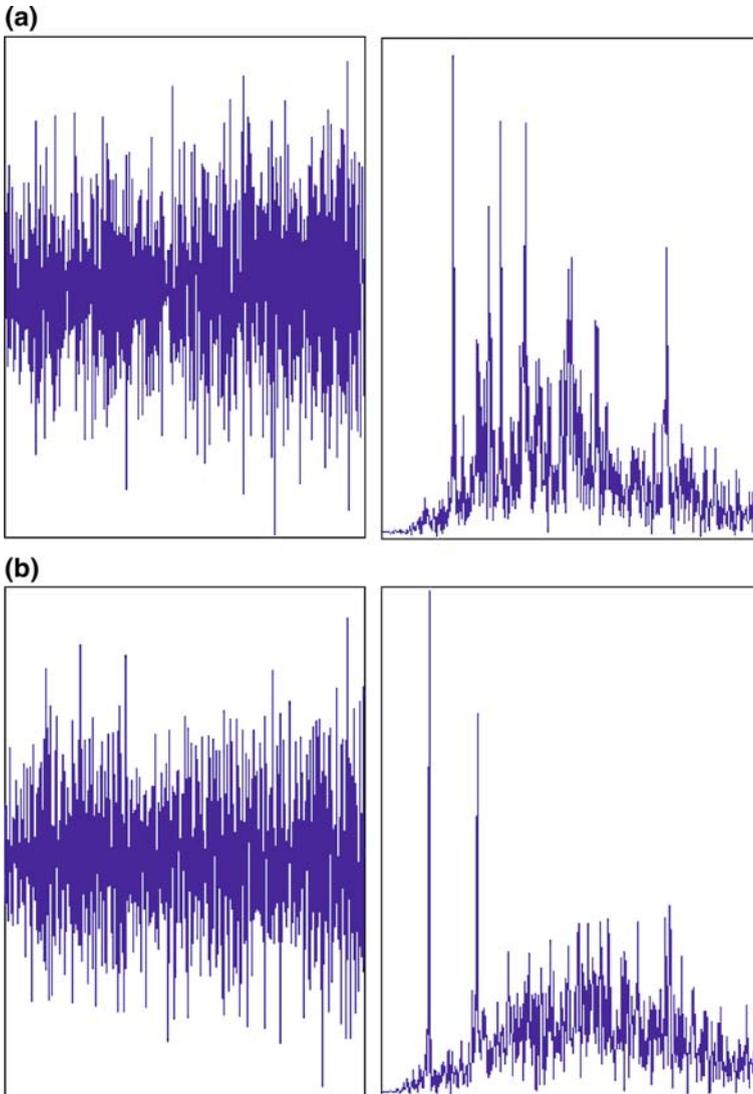


Fig. 2 Fragments of a truck and a van recordings and their spectra. Frames from left to right: truck, its spectrum; van, its spectrum

In order to enhance the robustness of the algorithm, we implement three different versions of RSNOFP that validate each other.

The sample signals for the training phase and the online signals in the detected phase are formed by imposing a comparatively short window on each input signal followed by a shift of this window along the signal so that adjacent windows have some overlap.

4.1 Outline of the approach

The complete process has three sequential steps:

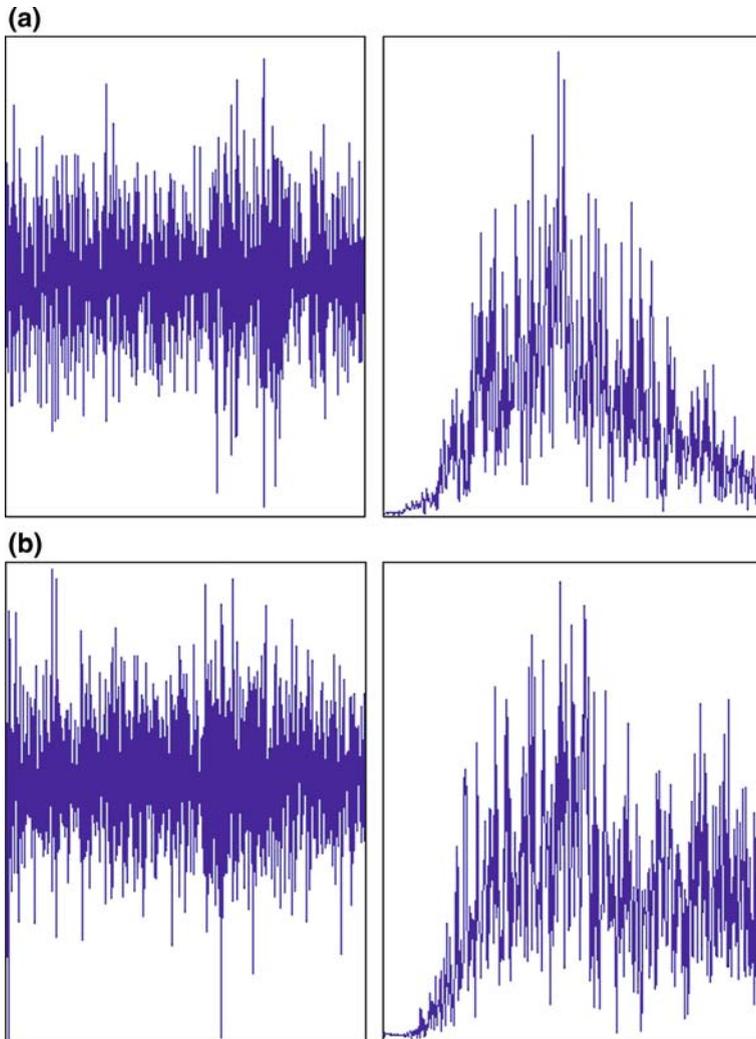


Fig. 3 Fragments of a plane and a helicopter recordings and their spectra. Frames from left to right: plane, its spectrum; helicopter, its spectrum

Training phase: We use a set of signals with known membership. These signals are sliced into the overlapped fragments of length L (typically, $L = 1024$). The fragments are subjected to the wavelet packet transform.

The wavelet packet transform bridges the gap between time domain and frequency domain representations of a signal. The coefficients from the upper levels (finer scales) correspond to basic waveforms, which are well localized in time domain but their spectra occupy wide frequency bands. The coefficients from the deeper levels (coarser scales) correspond to waveforms, which have narrow spectra but wide supports in time domain. We establish experimentally that the coefficients from the sixth level provide an optimal tradeoff between time and frequency domain representations of the 1024-samples acoustic fragments signals we are dealing with.

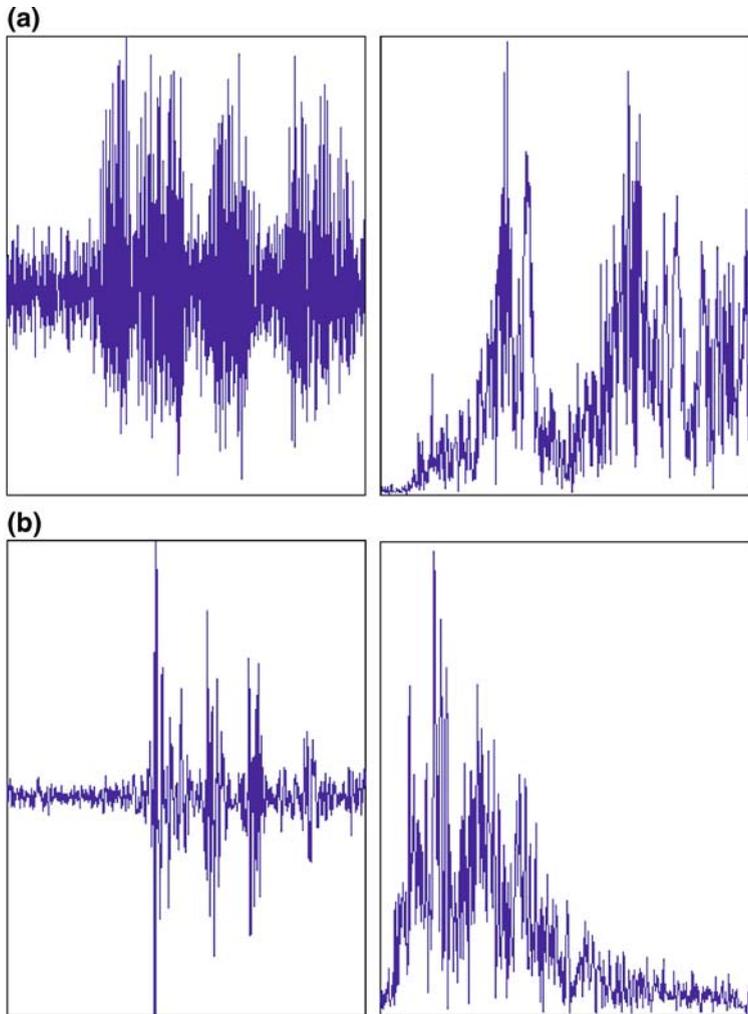


Fig. 4 Fragments of a speech and a wind recordings and their spectra. Frames from left to right: wind, its spectrum; speech, its spectrum

The blocks energies are calculated and three versions of RSNOFP are applied. As a result, we represent each fragment by three different vectors of length $l \ll L$ (typically, $l = 12$ or $l = 8$). The components of these vectors are the characteristic features of the fragments. These vectors are used as pattern data sets and also are utilized to produce three versions of CART trees.

Identification—features extraction phase: We slice the new acquired signal to overlapped fragments of length L . Then, the wavelet packet transform is applied followed by energies calculations of the blocks of coefficients. Then, we apply three different transforms that are determined by three versions of RSNOFP. As a result, we represent each fragment by the three different vectors of length l .

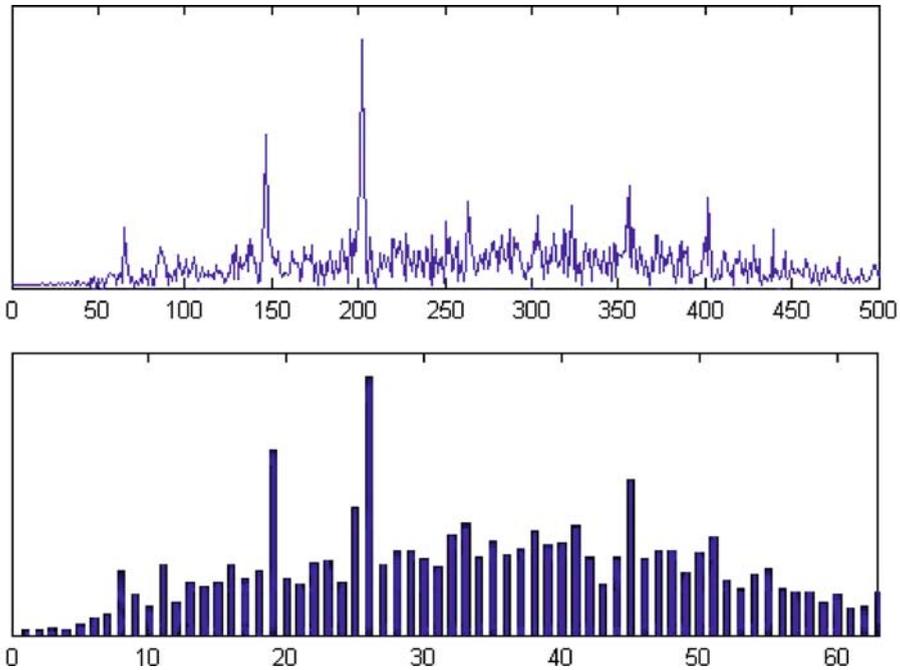


Fig. 5 Top: Fourier spectrum of the car signal in Fig. 1. Bottom: energies in the blocks of wavelet packet coefficients of the sixth level of the wavelet packet transform that uses the orthogonal spline wavelets of sixth order

Identification—decision making phase: These vectors are submitted to the corresponding versions of CART trees classifiers. In addition, the vectors are tested by a second classifier that calculates the distances of the vectors from the pattern data sets associated with V and N classes. The final decision on membership of the fragment is derived by combining the answers for all the three vectors from both classifiers.

5 Description of the algorithm and its implementation

5.1 The algorithm

The algorithm is centered around three basic phases:

- I. Extraction of characteristic features from V and N classes. It contains the following steps:
 1. The analyzing wavelet filters are chosen.
 2. The training sets of signals are constructed by slicing the training signals into overlapped segments.
 3. The wavelet packet transform is applied to segments.
 4. The energies in the blocks of the wavelet packet coefficients are calculated.
 5. RSNOFP is called. It results in the embedding of the training sets of signals into low-dimensional reference sets that contain its characteristic features.

- II. Building the CART classification trees.
- III. Identifying whether the new signal belongs to either V or N class:
 1. The new signal is sliced into overlapped segments.
 2. The wavelet packet transform is applied to these segments.
 3. The energies in the blocks of the wavelet packet coefficients are calculated.
 4. The set of blocks energies of each segment is embedded into a low-dimensional vector that contains its characteristic features.
 5. The distances of the vector that contains characteristic features from the reference sets of V and N classes are calculated.
 6. The vector is tested by CART classifier.
 7. Decision whether the vector belongs to either V or N class is made.

Now we present a detailed description of the implementation of this algorithm.

5.2 Implementation

5.2.1 Extraction of characteristic features

Choice of the analyzing waveforms: By now, a broad variety of orthogonal and biorthogonal filters, which generate wavelet packets coefficients, are available (Daubechies 1992; Mallat 1998; Averbuch and Zheludev 2007a,b). We use the sixth-order spline wavelet. This filter reduces the overlap between frequency bands associated with different decomposition blocks. At the same time, the transform with this filter provides a variety of waveforms that have a fair time-domain localization. For details see Appendix 1.

Signal preparation for training the algorithm: Initially, we gather as many recordings as possible for V and N classes, which have to be separated. Then, we prepare from each selected recording, which belongs to a certain class, a number of overlapped slices each of length $L = 2^J$ samples, shifted by $S \ll L$ samples with respect to each other. Altogether, we prepare M^v and M^n slices for the V and N classes, respectively. The slices are arranged to be rows of two matrices. Slices from V-signals form the matrix $A^v = \{A_{i,j}^v\}$, $i = 1, \dots, M^v$, $j = 1, \dots, L$ and slices from N-signals form the matrix $A^n = \{A_{i,j}^n\}$, $i = 1, \dots, M^n$, $j = 1, \dots, L$.

Embedding the sets of slices into sets of energies: First, the measure for the energy is specified. We use the normalized l_1 norms of the blocks. Then, the following operations are carried out:

1. The wavelet packet transform down to level m (typically $m = 6$ if $L = 1024$) is applied to each slice of length L from both V and N classes. We take the coefficients from the coarsest scale m . This scale contains $L = 2^J$ coefficients that are arranged into 2^m blocks of length $l = 2^{J-m}$, each of which is associated with a certain frequency band. These bands form a near-uniform partition of size 2^m of the Nyquist frequency domain.
2. The “energy” of each block is calculated using the chosen measure. We obtain, to some extent, the distribution of the “energies” of the slices $A^v(i, :)$ from V-signals and $A^n(i, :)$ from N-signals over various frequency bands of widths N_F/m , where N_F is the Nyquist frequency. The “energies” are stored in the energy vectors \bar{E}_i^v and \bar{E}_i^n of length $\lambda = 2^m = L/l$, respectively (typically, $\lambda = 64$). The energy vectors are arranged to be rows of two matrices $B^v = \{B_{i,j}^v\}$, $i = 1, \dots, M^v$, $j = 1, \dots, \lambda$ and $B^n = \{A_{i,j}^n\}$, $i = 1, \dots, M^n$, $j = 1, \dots, \lambda$, where the i th row of the matrix B^v is the vector \bar{E}_i^v and the i th row of the matrix B^n is the vector \bar{E}_i^n . The “energy” vectors $B^v(i, :)$ and

$B^n(i, :)$ are considered to be the averaged Fourier spectra of the slices $A^v(i, :)$ and $A^n(i, :)$, respectively, as it is seen in Fig. 5. We consider these vectors to be proxies of the slices. By the above operations we reduced the dimension of the database by factor $l = 2^{J-m}$.

Embedding of sets of energies into the sets of features: The subsequent operations yield a further reduction of the dimensionality in the compressed sensing (Donoho 2006; Donoho and Tsaig 2006) spirit. It is achieved by the application of three versions of the RSNOFP scheme to the energy matrices B^v and B^n . The RSNOFP scheme is described in Appendix 2. As a result, we get three pairs of the reference matrices: D_{rand}^v & D_{rand}^n , D_{pca}^v & D_{pca}^n , and D_{perm}^v & D_{perm}^n and the corresponding random matrices ρ_{rand} , ρ_{pca} , and ρ_{perm} . Rows of these matrices represent the slices of training signals. These random features matrices will be utilized in the identification phase.

Compaction of the feature matrices in V-class: In order to refine the V-class of feature matrices, we test their rows. Recall that each row is associated with a segment of a signal that belongs to V-class. We calculate the Mahalanobis distances d^v and d^n of each row in the matrix D_{rand}^v from the sets D_{rand}^v and D_{rand}^n . If for some row $d^v > d^n$, then, we remove this row from the matrix D_{rand}^v . The same is done for the matrices D_{pca}^v and D_{perm}^v .

Conclusion: As a result of the above operations, the dimensionality of the training set was substantially reduced. Typically, a segment of length 1024 is embedded into a 12-component vector. Ostensibly, this part of the process looks computationally expensive, especially if, for better robustness, large training sets are involved. This procedure is called once and it is done off-line before the detection phase that is done on-line. Altogether, formation of three pairs of reference matrices requires 2-3 min of CPU time on a standard PC.

Figure 6 displays one row from matrix D_{perm}^v and one row from matrix D_{perm}^n . These are sets of features from a segment in V-class and a segment in N-class.

5.3 Building the CARTs

Once we have three pairs of the reference matrices: D_{rand}^v & D_{rand}^n , D_{pca}^v & D_{pca}^n , and D_{perm}^v & D_{perm}^n , we proceed to build the classifiers. For this purpose we use vectors, which form rows in the reference matrices. The construction of the tree is done by a binary split of the space of input patterns $X \rightarrow \{X_1 \cup X_2 \cup \dots \cup X_r\}$, so that, once a vector appeared in the subspace X_k , its membership could be predicted with a sufficient reliability. The answer is the class the vector is assigned to and the probability of this assignment. The basic idea behind the split is that the data in each descendant subset is “purer” than the data in the parent subset. The scheme is described in full details in the monograph (Breiman et al. (1993)). A brief outline is given in Averbuch et al. (2001a).

After the construction of the three classification tree versions T_{rand} , T_{pca} , and T_{perm} with the three pairs of the reference matrices, we are in a position to classify new signals that were not used in the training phase.

5.4 Identification of an acoustic signal

An acoustic signal to be identified is preprocessed by the same operations that were used on the training signals.

Preprocessing of a new acoustic recording.

1. This recording is sliced to M overlapped segments of length $L = 2^J$ samples each, shifted with respect to each other by S samples. The wavelet packet transform up to scale m is applied to each slice. We take the coefficients from the coarsest scale m that are arranged into 2^m blocks of length $l = 2^{J-m}$. The “energy” of each block is calculated

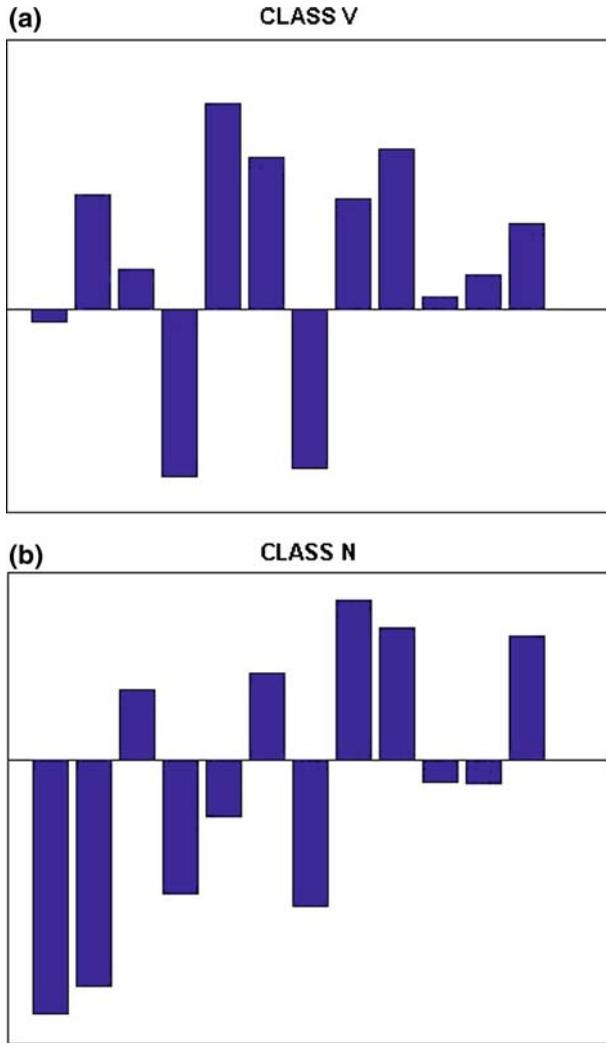


Fig. 6 Left: one row from the matrix D_{perm}^v (features of a segment from the V-class). Right: one row from the matrix D_{perm}^n (features of a segment from the N-class)

with the chosen measure. Thus, each i th slice is embedded into an energy vector \vec{E}_i of length $\lambda = 2^m = L/l$. The energy vectors are arranged in the matrix $B = \{B_{i,j}\}$, $i = 1, \dots, M$, $j = 1, \dots, \lambda$, where the i th row of the matrix B is the vector \vec{E}_i .

- In order to embed the energy matrix B into the features spaces, we multiply it subsequently by the random matrices ρ_{rand} , ρ_{pca} , and ρ_{perm} . These multiplications produce three features matrices: D_{rand} , D_{pca} , and D_{perm} , where the i th row in each matrix is associated with the i th segment of the processed signal.

Identification of a single segment. To identify the i th segment of a signal, we take three vectors, \vec{v}_{rand}^i , \vec{v}_{pca}^i , and \vec{v}_{perm}^i , which form the i th rows of the matrices D_{rand} , D_{pca} , and D_{perm} , respectively.

1. These vectors are submitted to the respective versions T_{rand} , T_{pca} , and T_{perm} of the classification tree. Once a vector is submitted to the tree, it is assigned to one of the subsets X_k of the input space X . These trees produce three decisions τ_{rand} , τ_{pca} , and τ_{perm} together with the corresponding probabilities p_{rand} , p_{pca} , and p_{perm} . The decision $\tau_{(\cdot)}$ determines the most probable membership of the segment. The value $\tau_{(\cdot)} = 1$ if the CART assigns the segment to V-class and $\tau_{(\cdot)} = 0$ otherwise.
2. The distances (for example, Mahalanobis or Euclidean) between the vectors \vec{v}_{rand}^i , \vec{v}_{pca}^i , and \vec{v}_{perm}^i and the respective pairs of the reference sets: D_{rand}^v & D_{rand}^n , D_{pca}^v & D_{pca}^n , and D_{perm}^v & D_{perm}^n are calculated. This calculation produces three decisions $\tilde{\tau}_{rand}$, $\tilde{\tau}_{pca}$, and $\tilde{\tau}_{perm}$ together with the corresponding quasi-probabilities (QP) \tilde{p}_{rand} , \tilde{p}_{pca} , and \tilde{p}_{perm} in the following way. Let d^v and d^n be the distances of a vector $\vec{v}_{(\cdot)}^i$ from the respective pair of the reference sets: $D_{(\cdot)}^v$ and $D_{(\cdot)}^n$. If $d^v < d^n$ then the decision is $\tilde{\tau}_{(\cdot)} = 1$ (the segment is assigned to the V-class), otherwise $\tilde{\tau}_{(\cdot)} = 0$ (the segment is assigned to the N-class). If $d^v < d^n$ then the membership QP in the V-class is defined as $\tilde{p}_{(\cdot)} = 1 - d^v/d^n$, otherwise $\tilde{p}_{(\cdot)} = 0$. This classification scheme is a version to the Nearest Neighbor scheme. We call this scheme the Minimal Distance (MD) classifier.
3. Two thresholds values t and \tilde{t} are set and the results for the i th segment are combined into three 3-component column vectors \vec{y}_{rand}^i , \vec{y}_{pca}^i , and \vec{y}_{perm}^i , where:

$$\begin{aligned}
 y_{(\cdot)}^i(1) &= \begin{cases} 1, & \text{if } p_{(\cdot)} > t \\ 0, & \text{otherwise,} \end{cases} \\
 y_{(\cdot)}^i(2) &= \begin{cases} 1, & \text{if } \tilde{p}_{(\cdot)} > \tilde{t} \\ 0, & \text{otherwise,} \end{cases} \\
 y_{(\cdot)}^i(3) &= \tau_{(\cdot)} \times \tilde{\tau}_{(\cdot)}.
 \end{aligned}
 \tag{1}$$

Identification of a recording.

1. The vectors \vec{y}_{rand}^i , \vec{y}_{pca}^i , and \vec{y}_{perm}^i are gathered into three matrices Y_{rand} , Y_{pca} , and Y_{perm} of size $3 \times M$, where M is the number of overlapping segments produced from the analyzed signal. The vectors $\vec{y}_{(\cdot)}^i$ serve as the i th columns in the respective matrices Y .
2. The rows of the matrices are processed by moving average.
3. The matrices Y_{rand} , Y_{pca} , and Y_{perm} are combined into the matrix Y in the following way. Each entry in Y is defined as the median value of the respective entries of the three matrices:

$$Y(i, j) = \text{median} (Y_{rand}(i, j), Y_{pca}(i, j), Y_{perm}(i, j)). \tag{2}$$

Conclusions. The matrix Y contains the results for the analyzed signal. Its first row contains the averaged answers (which have a significant probabilities) from the CART classifier. Its value at each point is the number of positive (class V) answers in the vicinity of this point, which is divided by the length of the vicinity. It represents the “density” of the positive answers around the corresponding segment. The structure of the second row is similar to the structure of the first row with the difference that these answers come from the MD classifier instead of answers from the CART classifier. The third row of the matrix Y combines the answers from both classifiers. First, these answers are multiplied by each other. The combined answer is equal to one for segments where both classifiers produce answer one (V-class) and zero otherwise. Thus, the classifiers cross-validate each other. Then, the results are processed by the application of the moving average providing the “density” for the positive answers. This row in the matrix Y , yields the most robust result from the detection process with minimal false alarm.

Real-time implementation. Above, we described an identification process of a full recording. However, a slight modification of the scheme provides a real-time identification of an acoustic signal. To get it, we process separately each segment upon its arrival (of course, there segments are overlapped). This segment-wise processing takes a fraction of a second. Once several initial segments are identified, we start a moving average procedure. Thus, the arriving signal is identified with a negligible time delay.

We presented a scheme for the detection of the arrival of any moving vehicle. Obviously, the scheme is also applicable for the detection of the arrival of the sought after vehicles.

6 Experimental results

We conducted a series of experiments to detect the arrival of vehicles of arbitrary type in the presence of surrounding noises.

Altogether 200 recordings were available taken in five different areas. Many recordings contained sounds emitted by different vehicles combined with the sounds of wind, speech, aircrafts, etc. The original SR was 48,000 SPS. The signals were downsampled into SR of 1,000 SPS. The motion dynamics, the distances of vehicles from the receiver, the surrounding conditions were highly diverse.

6.1 Detection experiments

The first task was to form the reference database of signals with known membership (training) for building the classifiers. This database was derived mainly from the recordings made on a single date by clipping the corresponding fragments. A few recordings from other dates were also used. The CAR fragments were extracted from ten recordings, five recordings were used for the TRUCK fragments and the same number for the VAN fragments. The diverse non-vehicle fragments were extracted from 35 recordings. Altogether, 38 recordings were involved in the training process (most of them contained sounds from different sources). We tested various families of wavelet packets and various norms for the feature extraction and various combinations of features presented to the MD and CART classifiers. The best results were achieved with wavelet packet transform that uses the sixth-order spline filters and the l_1 norm.

We separated the reference signals into two groups. One group (V class) contains all signals associated with vehicles and the other group (N class) contains all the non-vehicles signals. The signals were sliced into overlapped segments of length $L = 1,024$ that were shifted with respect to one another by $S = 256$, thus, the overlap was $3/4$ of a fragment. Altogether, we obtained 21,898 V-segments and 43,816 N-segments. We extracted the characteristic features from the segments as explained in Sect. 5.2.1. Each segment was expanded by the wavelet packet transform down to sixth level (scale) and the l_1 norm was used as the “energy” measure for all the 64 blocks of the sixth level. As a result of the procedures that were described in Sect. 5.2.1, we selected various sets of discriminant blocks. These procedures produced three pairs of reference matrices: D_{rand}^v & D_{rand}^n , D_{pca}^v & D_{pca}^n , and D_{perm}^v & D_{perm}^n with the corresponding random matrices ρ_{rand} , ρ_{pca} , and ρ_{perm} . Each matrix has 12 columns according to the number of the characteristic features. These matrices were used for the MD classification and also were utilized for building three CART trees T_{rand} , T_{pca} , and T_{perm} . For the MD classification, we used all the available features (12), unlike building the CART trees, where better results were achieved with sets containing only 8 features.

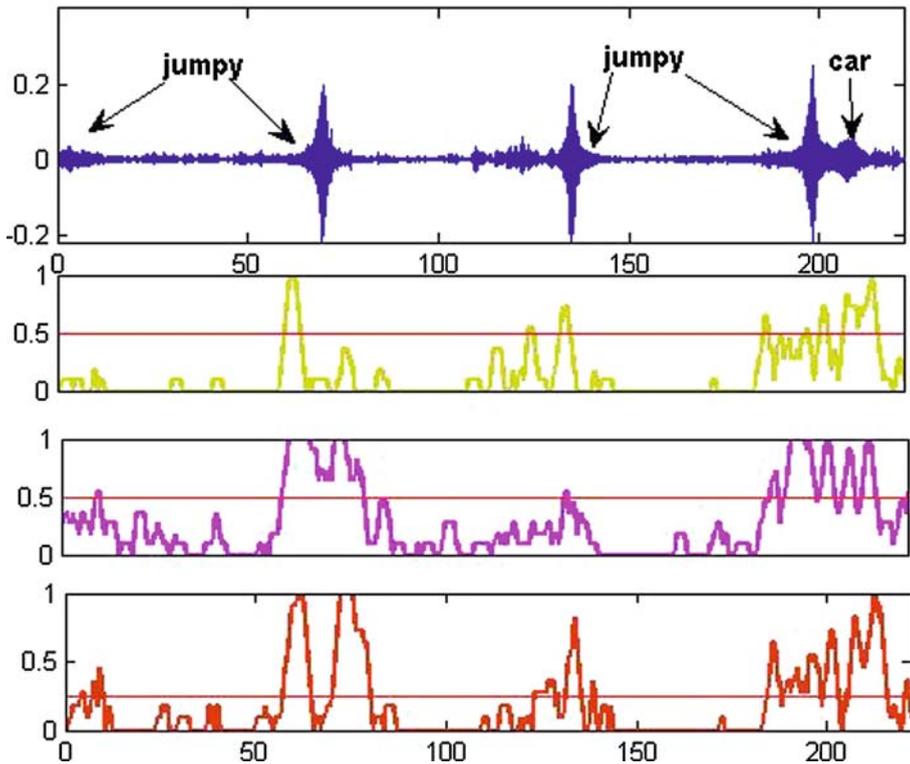


Fig. 7 Results of testing the recording # 1. In the beginning the sound from a remote vehicle is heard. The jumpy vehicle passed by the receiver at around 70 s from the start, at around 134 s, and at around 200 s, in last passage it was followed by a car

All of the 200 available recordings were employed in the detection phase. A recording number k was embedded into three features matrices D_{rand}^k , D_{pca}^k , and D_{perm}^k , where the i th row of each matrix is associated with the i th segment of the recording (see Sect. 5.4). Each row was tested with the MD and CART classifiers. The results were gathered into the Y^k matrix. The Euclidean distance was used for the MD classification. Although the signals were recorded in very different conditions and many of them were corrupted by strong background noise, the arrival of vehicles was correctly detected in overwhelming majority of cases, while false alarm was minimal.

In Figs. 7–12 we present a few typical results from the experiments on detection of vehicles of arbitrary types. All the figures are identically organized. Each comprises four frames. The top frame depicts the original recording # k . The next three frames present five rows from the Y^k matrix with respect to time scale. The second from the top frame presents the combined answers (the median from three answers) from CART classifiers processed by the moving average. The third from the top frame similarly displays the results from the MD classifiers. The bottom frame illustrates the combined results from both classifiers, that is the answers from both classifiers multiplied with each other and processed by the moving average.

All the displayed recordings did not participate in the training phase.

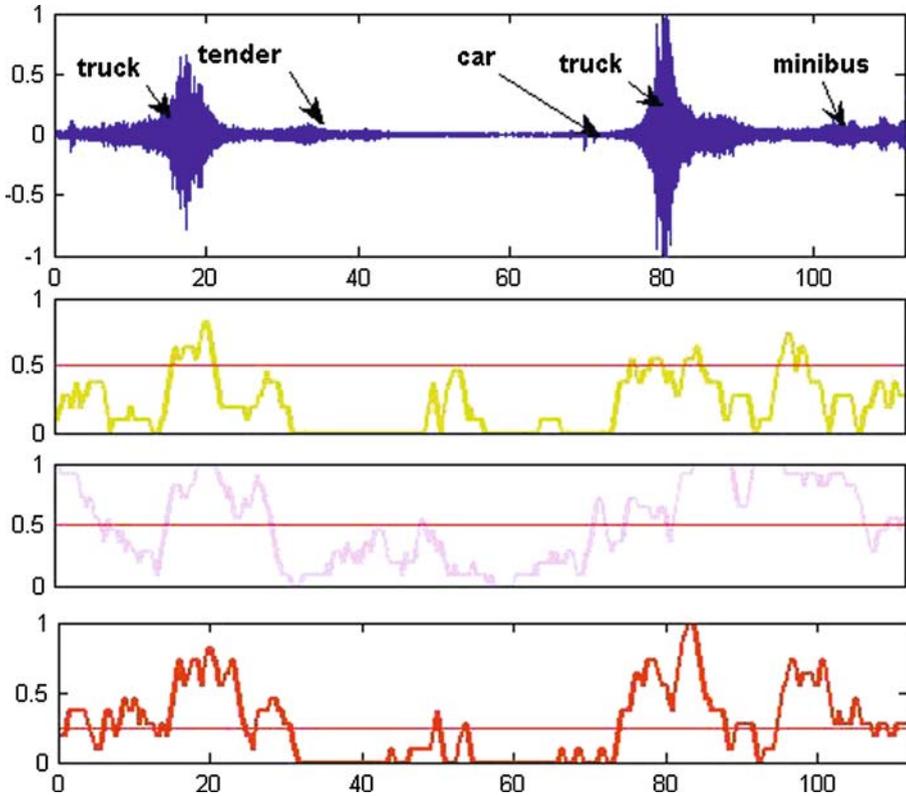


Fig. 8 Results of testing the recording # 2. In the beginning of the recording a truck passed by the receiver followed by a tender. At around 70s from the start, a car followed by a truck passed by. In the end a minibus and a car arrived. The recording did not participate in the training phase

6.1.1 Examples

Recording # 1: We display in Fig. 7 the results from testing recordings # 1. In the beginning of the recording, sound from a remote vehicle is heard. Then, the jumpy vehicle passed by the receiver at around 70, 134, and 200s from start. In its last passage it was followed by another car. All the events were correctly detected by the CART and MD classifiers. The MD classifier produced some false alarms, which were eliminated by combining the classifiers (bottom frame).

Recording # 2: We display in Fig. 8 the results from testing recordings # 2. In the beginning of the recording, a truck passed by the receiver followed by a tender. At around 70s from the start of the recording a car followed by a truck passed by. In the end, a minibus and a car arrived. All the events were correctly detected by the CART and MD classifiers for each SR. The MD classifier produced some false alarms, which were reduced by combining the classifiers (bottom frame).

Recording # 3: We display in Fig. 9 the results from testing recordings # 3. Two trucks passed by the receiver moving in opposed directions at around 50s from the start. Strong wind is present. The event was correctly detected by the CART and MD classifiers for each SR. The MD classifier produced some false alarms, which were reduced by combining the classifiers (bottom frame).

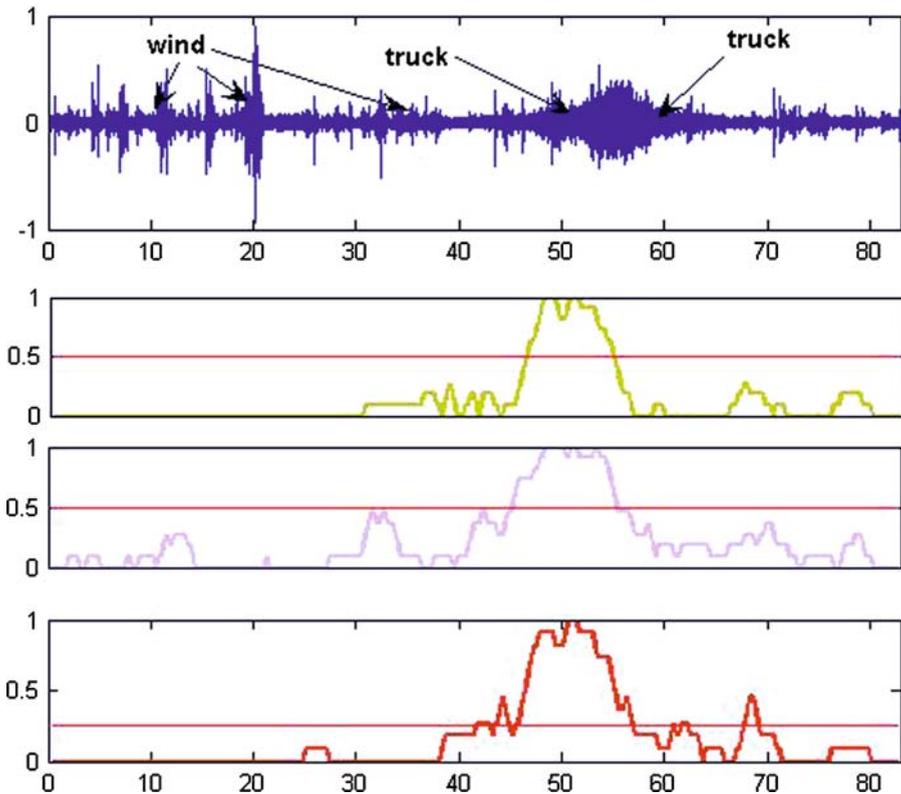


Fig. 9 Results of testing the recording # 3. Two trucks passed by the receiver in opposed directions at around 50 s from the start, the strong wind presented at the scene

Recording # 4: We display in Fig. 10 the results from testing recordings # 4. A truck passed by the receiver from 30 to 190 s from the start. Then, a strong sound from a plane dominated the scene till the end of the recording. The truck was correctly detected by the CART and MD classifiers. The MD classifier produced some false alarms, which were reduced by combining the classifiers (bottom frame). The plane was not assigned to the V class.

Recording # 5: We display in Fig. 11 the results from testing recordings # 5. A truck followed by a minibus passed by the receiver at around 40 s from the start and one more truck at around 65 s. Strong wind was present. The vehicles were correctly detected by the CART and MD classifiers. The MD classifier produced some false alarms, which were eliminated by combining the classifiers (bottom frame).

Recording # 6: We display in Fig. 12 the results from testing recordings # 6. A sound from a truck was heard within the intervals of 15–50 and 80–110 s from the start. Then, a plane sound appeared. It lasted till the end of the recording. The truck was correctly detected by the CART and MD classifiers. The plane was not assigned to V class. The MD classifier performed worse.

Comments

- The detection experiments demonstrate the relevance of our approach to feature extraction.

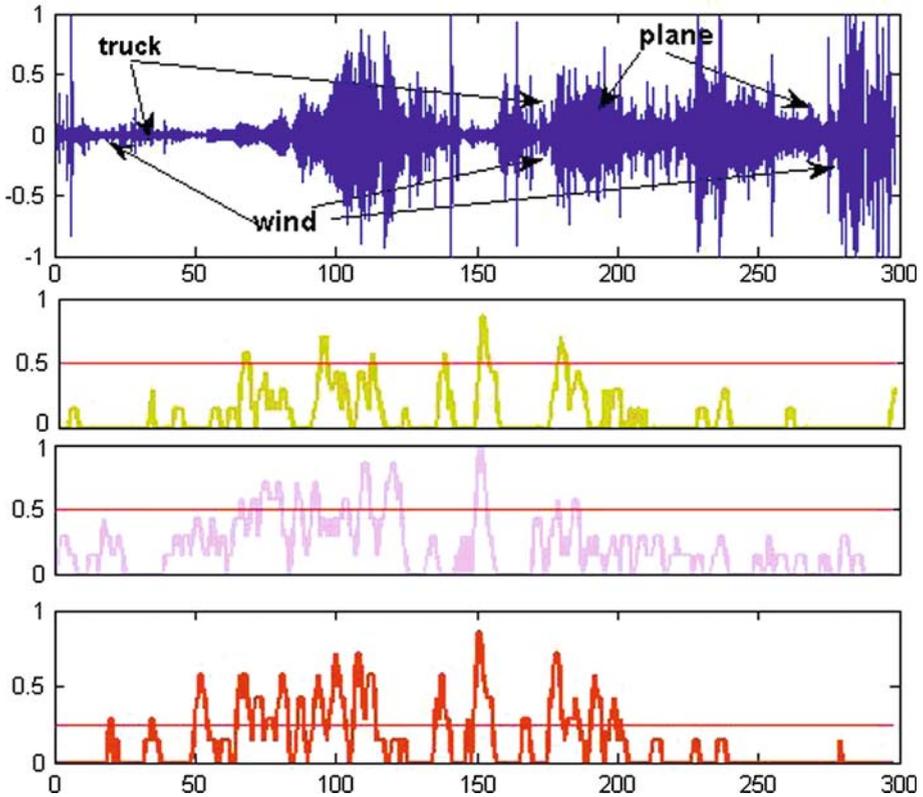


Fig. 10 Results of testing the recording # 4. A truck passed by the receiver from 30 to 190s from the start, then the strong sound from a plane dominated the scene till the end of the recording. Strong wind presented

- Combination of three schemes for feature extraction performs better than any single scheme.
- Combining the classifiers MD with CART significantly reduces the number of false alarms.
- The algorithm produced satisfactory detection results even when the conditions of the real signals essentially differs from the training data and the surrounding conditions. When the conditions of the captured signals were close to the training conditions, the detection is almost perfect.
- The algorithm performs similarly on signal with sampling rates of 1,000 and 600 SPS. In few cases, the results for SR of 1,000 SPS were significantly better than those for SR of 600 SPS.

7 Conclusions

We presented a robust algorithm that detects the arrival of a vehicle of arbitrary type via the analysis of its acoustic signature against an existing database of recorded and processed acoustic signals.

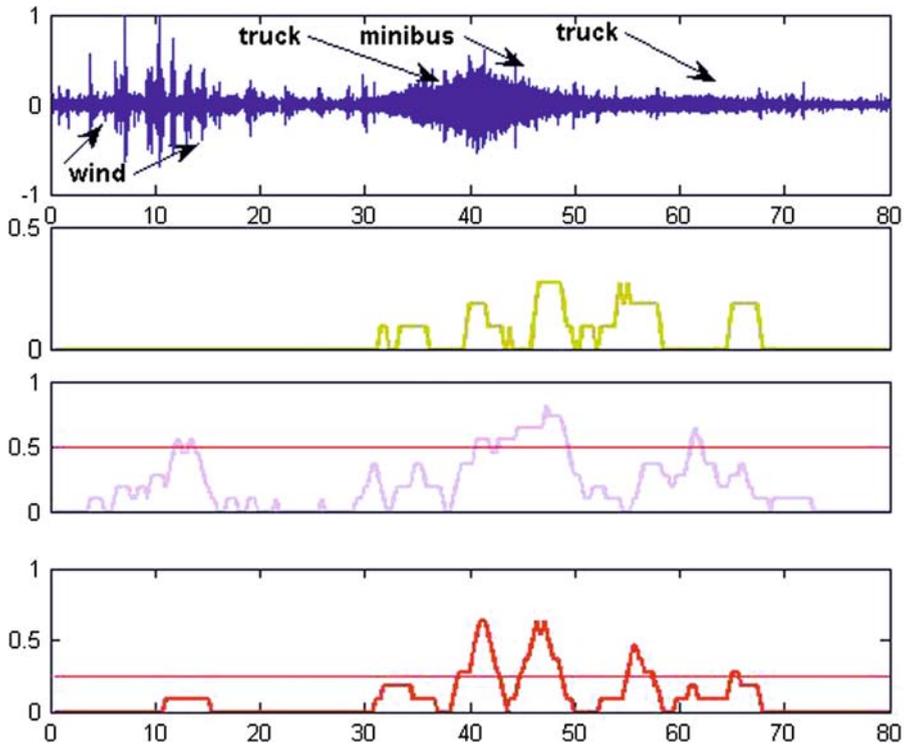


Fig. 11 Results of testing the recording # 5. A truck followed by a minibus passed by the receiver at around 40s from the start and one more truck at around 65 s. The strong wind presented at the scene

To minimize the number of false alarms, we constructed an acoustic signature of a certain vehicle using the distribution of the energies among blocks which consist of its wavelet packet coefficients. This distribution serves as an averaged version of the Fourier spectrum of the signal. To reduce the dimensionality of the features sets, we designed the scheme of RSNOFP, which proved to be an efficient tool for the extraction of a small number of characteristic features of the objects to be detected.

As decision units for detection, a classifier that is based on the MD from the reference data sets and the CART classifier were used. These classifiers cross-validated each other.

The detection process is fast and can be implemented in real time.

This technology, which has many algorithmic variations, is generic and can be used to solve a wide range of classification and detection problems such as process control, which are based on acoustic processing and, more generally, for classification and detection of signals which have near-periodic structure. Distinguishing between different vehicles can also be achieved via this technology.

Appendix 1: wavelet and wavelet packet transforms

By now the wavelet and wavelet packet transforms are widespread and have been described comprehensively in the literature (Daubechies 1992; Wickerhauser 1994; Mallat 1998).

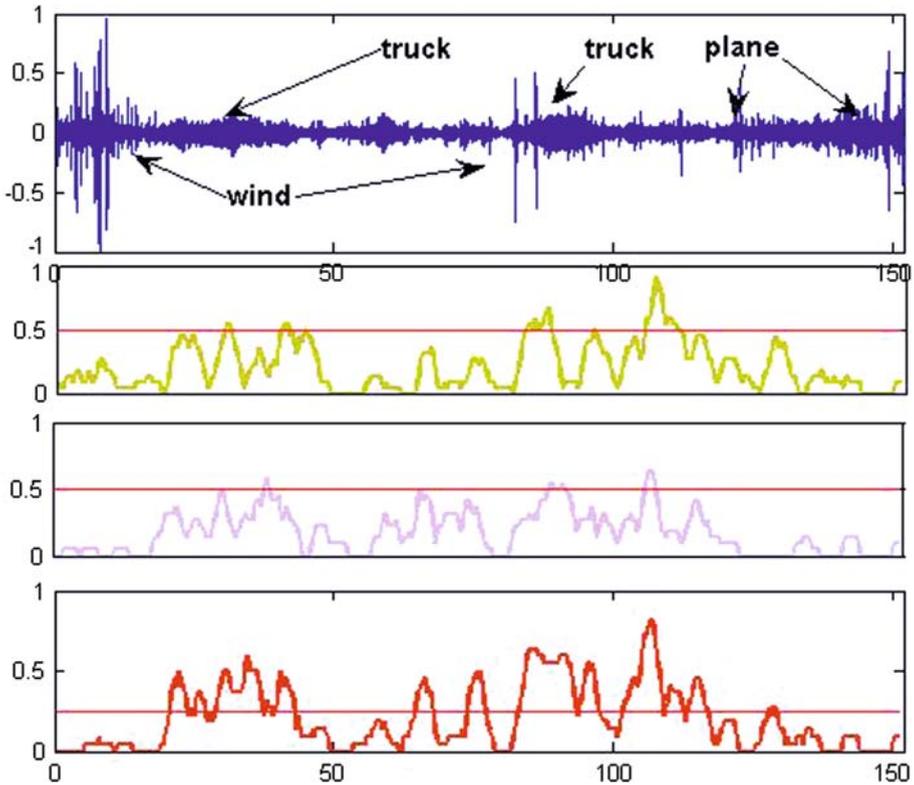


Fig. 12 Results of testing the recording # 6. The sound from a truck was heard within the intervals 15–50 and 80–110 s from the start, then the sound from a plane appeared at the scene, which lasted till the end of the recording

Therefore, we restrict ourselves to mention only relevant facts that are necessary to understand the construction of the algorithm.

The result from the application of the wavelet transform to a signal f of length $n = 2^J$ is a set of n correlated coefficients of the signal with scaled and shifted versions of two basic waveforms—the father and mother wavelets. The transform is implemented through iterated application of a conjugate pair of low- (H) and high- (G) pass filters followed by downsampling. In the first decomposition step, the filters are applied to f and, after downsampling, the result has two blocks w_0^1 and w_1^1 of the first scale coefficients, each of size $n/2$. These blocks consist of the correlation coefficients of the signal with 2-sample shifts of the low frequency father wavelet and high frequency mother wavelet, respectively. The block w_0^1 contains the coefficients necessary for the reconstruction of the low-frequency component of the signal. Because of the orthogonality of the filters, the energy (l_2 norm) of the block w_0^1 is equal to that of the component W_0^1 . Similarly, the high frequency component W_1^1 can be reconstructed from the block w_1^1 . In this sense, each decomposition block is linked to a certain half of the frequency domain of the signal.

While block w_1^1 is stored, the same procedure is applied to block w_0^1 in order to generate the second level (scale) of blocks w_0^2 and w_1^2 of size $n/4$. These blocks consist of the correlation coefficients with 4-sample shifts of the two times dilated versions of the father and mother wavelets. Their spectra share the low frequency band previously

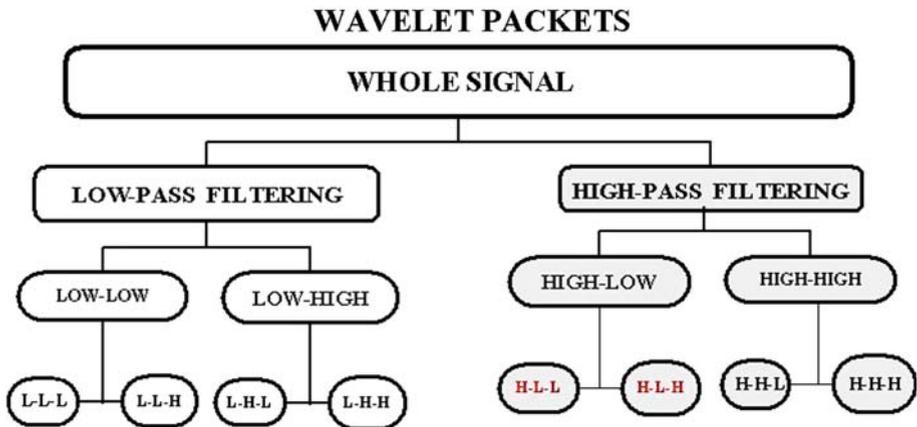


Fig. 13 Flow of the wavelet packet transform

occupied by the original father wavelet. Then, w_0^2 is decomposed in the same way and the procedure is repeated m times. Finally, the signal f is transformed into a set of blocks $f \rightarrow \{w_0^m, w_1^m, w_1^{m-1}, w_1^{m-2}, \dots, w_1^2, w_1^1\}$ up to the m th decomposition level. This transform is orthogonal. One block is remained at each level (scale) except for the last one. Each block is related to a single waveform. Thus, the total number of waveforms involved in the transform is $m + 1$. Their spectra cover the whole frequency domain and split it in a logarithmic form. Each decomposition block is linked to a certain frequency band (not sharp) and, since the transform is orthogonal, the l_2 norm of the coefficients of the block is equal to the l_2 norm of the component of the signal f whose spectrum occupies this band.

Through the application of the wavelet packet transform, many more waveforms, namely, 2^j waveforms at the the j th decomposition level are involved. The difference between the wavelet packet and wavelet transforms begins in the second step of the decomposition. Now both blocks w_0^1 and w_1^1 are stored at the first level and at the same time both are processed by pair of H and G filters which generate four blocks $w_0^2, w_1^2, w_2^2, w_3^2$ in the second level. These are the correlation coefficients of the signal with 4-sample shifts of the four libraries of waveforms whose spectra split the frequency domain into four parts. All of these blocks are stored in the second level and transformed into eight blocks in the third level, etc. The involved waveforms are well localized in time and frequency domains. Their spectra form a refined partition of the frequency domain (into 2^j parts in scale j). Correspondingly, each block of the wavelet packet transform describes a certain frequency band.

Flow of the wavelet packet transform is given by Fig. 13. The partition of the frequency domain corresponds approximately to the location of blocks in the diagram.

There are many wavelet packet libraries. They differ from each other by their generating filters H and G , the shape of the basic waveforms and their frequency content. In Fig. 14 we display the wavelet packets after decomposition into three levels generated by the spline of sixth order. While the splines do not have a compact support in time domain, they are localized fairly. They produce perfect splitting of the frequency domain.

There is a duality in the nature of the wavelet coefficients of a certain block. On one hand, they indicate the presence of the corresponding waveform in the signal and measure its contribution. On the other hand, they evaluate the contents of the signal inside the related frequency band. We may argue that the wavelet packet transform bridges the gap between

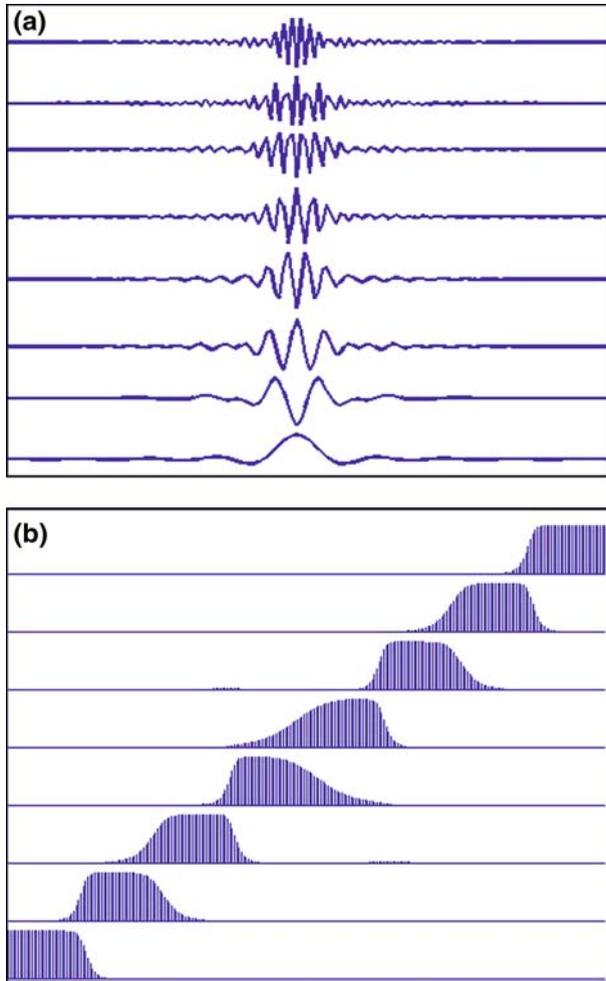


Fig. 14 Spline-6 wavelet packets of the third scale (left) and their spectra (right)

time-domain and frequency-domain representations of a signal. As we advance to coarser level (scale), we see a better frequency resolution at the expense of time domain resolution and vice versa. In principle, the transform of a signal of length $n = 2^J$ can be implemented up to the J th decomposition level. At this level there exist n different waveforms, which are close to the sine and cosine waves with multiple frequencies. In Fig. 15, we display a few wavelet packets after decomposition into six levels generated by spline of the sixth order. The waveforms resemble the windowed sine and cosine waves, whereas their spectra split the Nyquist frequency domain into 64 bands.

Appendix 2: RSNOFP scheme

RSNOFP: version I. A random matrix R_1 of size $r \times \lambda$ where $r \ll \lambda$ (typically, $r = 20$) is created. Entries of the matrix R_1 are Gaussian random variables. The rows in the matrix are

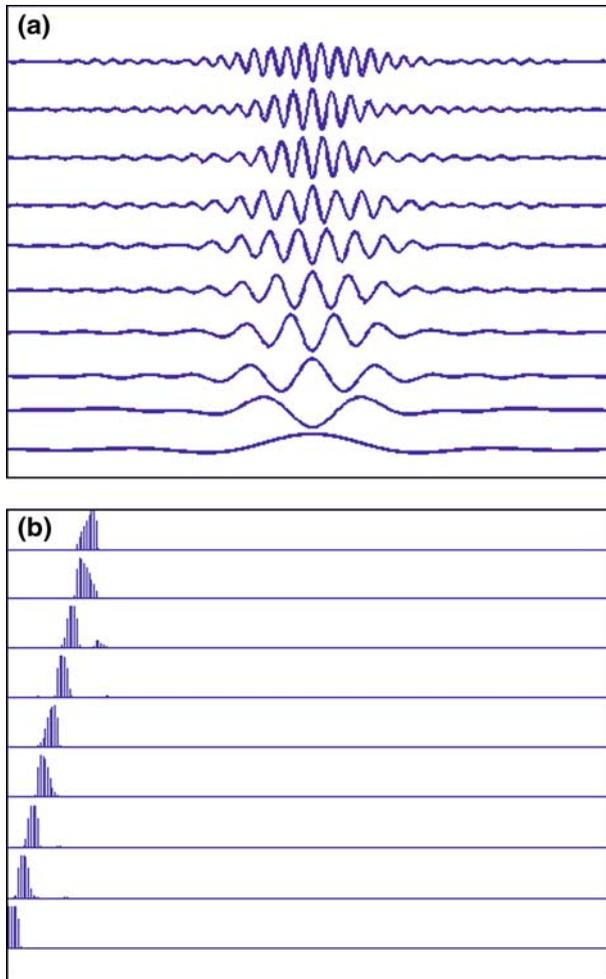


Fig. 15 Spline-6 wavelet packets of sixth scale (left) and their spectra (right)

normalized. The matrix B^v (defined in Sect. 5.2) is multiplied by the matrix R_1 . As a result, we obtain a new matrix $C^v = B^v \cdot R_1 = \{C_{i,j}^v\}$ of size $M^v \times r$. Each row in C^v is associated with the corresponding slice from A^v . To select the most valuable columns in the matrix C^v , we average this matrix along the columns

$$\bar{c}^v = \frac{1}{M^v} \sum_{i=1}^{M^v} |C_{i,j}^v| = \{c_j^v\}, \quad j = 1, \dots, r.$$

Let K be the set of indices $k < r$ of the largest coordinates of the vector \bar{c}^v (typically, $k = 12$). Then, we remove the columns, whose indices do not belong to K , from the matrix C^v and obtain the matrix D^v of size $M^v \times k$. This operation is equivalent to multiplication of B^v with the matrix ρ of size $k \times \lambda$, which is derived from R_1 by removing the rows, whose indices do not belong to K . Thus, the initial matrix A^v consisting of the V-class slices, whose size

was, for example, $M^v \times 1,024$, is reduced to the matrix D^v of the *random footprints* of the slices. The size of D^v is $M^v \times 12$. To produce a similar reduction for the matrix A^n in N-class slices, we multiply the N-class energy matrix B^n with the matrix ρ . As a result, we obtain the random footprints matrix $D^n = B^n \cdot \rho$ of size $M^n \times 12$. We consider the coordinates of the i th row of the matrix $D^{v(n)}$ as the set of k characteristic features of the i th slice from the matrix $A^{v(n)}$.

Now, the Mahalanobis distances $\mu_i, i = 1, \dots, M^v$ of each row in the V-class matrix D^v from the matrix D^n are computed. Then, the sequence $\{\mu_i\}$ is averaged

$$\Delta = \frac{1}{M^v} \sum_{i=1}^{M^v} \mu_i.$$

The value Δ is considered to be distance between the sets of features D^v and D^n . The matrices D^v, D^n, ρ , and the value Δ are stored and we proceed to optimize the features.

All the above operations are conducted using a random matrix R_2 , whose structure is similar to the structure of the matrix R_1 . As a result, we obtain the features matrices D_2^v and D_2^n , the random matrix ρ_2 and the distance value Δ_2 . The distance value Δ_2 is compared to the stored value Δ . Assume, $\Delta_2 > \Delta$. This means that the features matrices D_2^v and D_2^n are better separated from each other than the stored matrices D^v and D^n . In this case, we denote D_2^v, D_2^n, ρ_2 , and the value Δ_2 as D^v, D^n, ρ , and the value Δ , respectively. They are stored while replacing the previous stored items. If $\Delta_2 \leq \Delta$ then the stored items are left intact.

We iterate this procedures up to 500 times. In the end, we stored the features matrices D^v and D^n such that the “distance” Δ between them among all the iterations is maximal. We have stored the reduced random matrix ρ and the pattern matrices D^v, D^n , which will be used in the identification phase. These items are denoted as D_{rand}^v, D_{rand}^n , and ρ_{rand} .

RSNOFP: Version II. This version is similar, to some extent, to Version I. The difference is that, instead of selecting the most valuable columns in the matrix C^v of size $M^v \times r$, we apply to this matrix the principal component analysis. As a result, we obtain the matrix $P = \{P_{i,j}\}$ of size $r \times r$. Each column of P contains coefficients for one principal component. The columns are arranged in decreasing component variance order. The size of P is reduced to $r \times k$ by retaining only the first k columns

$$P_k = \{P_{i,j}\}, \quad i = 1, \dots, r, \quad j = 1, \dots, k.$$

We obtain the feature matrix D^v for the V-class by multiplying C^v by P_k :

$$D^v = C^v \cdot P_k = B^v \cdot R_1 \cdot P_k = B^v \cdot \rho, \quad \text{where } \rho = R_1 \cdot P_k.$$

The size of the matrix ρ is $k \times \lambda$. Similarly, we produce the feature matrix D^n for the N-class: $D^n = B^n \cdot \rho$. Similarly to Version I, we measure the “distance” Δ between the feature sets D^v and D^n . The matrices D^v, D^n, ρ , and the value Δ are stored and we proceed to optimize the features, which is identical Version I. In the end, we stored the features matrices D^v and D^n and the matrix ρ . These items are denoted by D_{pca}^v, D_{pca}^n , and ρ_{pca} .

RSNOFP: version III. This version differs from versions I and II. Here, we do not multiply the energy matrix B^v by a random matrix but instead we perform a random permutation of the columns and retain the first r columns. Thus, we get the matrix C^v of size $M^v \times r$. Note, that this transform can be presented as the multiplication of the matrix B^v by a matrix T of size $\lambda \times r, C^v = B^v \cdot T$ where each column consists of zeros except for one entry, which is equal to 1.

Example: Assume that the matrix T is of size 4×3 that executes the permutation $[1\ 2\ 3\ 4] \rightarrow [3\ 1\ 4\ 2]$ of the columns of a matrix of size 4×4 while retaining the three first columns:

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The other operations are similar to the operations in Version II. We apply to the matrix C^v the PCA algorithm, which results in the matrix $P = \{P_{i,j}\}$ of size $r \times r$ of coefficients of principal components. The size of P is reduced to $r \times k$ by retaining only the first k columns,

$$P_k = \{P_{i,j}\}, \quad i = 1, \dots, r, \quad j = 1, \dots, k.$$

We obtain the feature matrix D^v for the V-class by multiplying C^v by P_k :

$$D^v = C^v \cdot P_k = B^v \cdot R_1 \cdot P_k = B^v \cdot \rho, \quad \text{where } \rho = R_1 \cdot P_k.$$

The size of the matrix ρ is $k \times \lambda$. Similarly, we produce the feature matrix D^n for the N-class: $D^n = B^n \cdot \rho$. We measure the “distance” Δ between the sets of features D^v and D^n . The matrices D^v , D^n , ρ , and the value Δ are stored and we proceed to optimize the features, which is identical to Versions I and II. In the end, the features matrices D^v and D^n and the matrix ρ are stored. These items are denoted by D^v_{perm} , D^n_{perm} , and ρ_{perm} .

We illustrate the RSNOFP procedures (version II) by the diagram in Fig. 16.

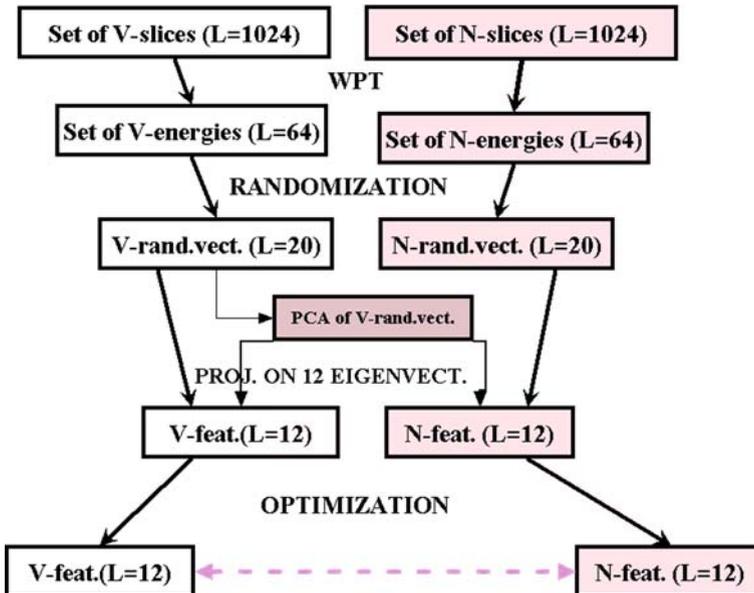


Fig. 16 RSNOFP procedures (version II)

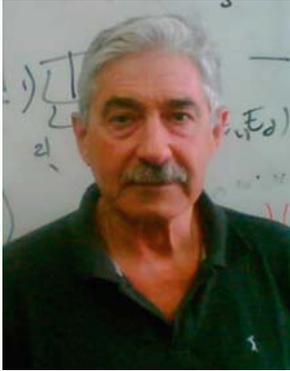
References

- Averbuch, A. Z., Hulata, E., Zheludev, V. A., & Kozlov, I. (2001a). A wavelet packet algorithm for classification and detection of moving vehicles. *Multidimensional Systems and Signal Processing*, 12(1), 9–31.
- Averbuch, A., Kozlov, I., & Zheludev, V. (2001b). Wavelet packet based algorithm for identification of quasi-periodic signals. In A. Aldroubi, A. F. Laine, & M. A. Unser (Eds.), *Proc. SPIE 4478, Wavelet Applications in Signal and Image Processing IX* (pp. 353–360).
- Averbuch, A., & Zheludev, V. (2007a). Wavelet transforms generated by splines. *International Journal of Wavelets, Multiresolution and Information Processing*, 5(2), 257–292.
- Averbuch, A., & Zheludev, V. (2007b). Wavelet and frame transforms originated from continuous and discrete splines. In J. Astola & L. Yaroslavsky (Eds.), *Advances in signal transforms: Theory and applications* (Chapter 1, pp. 1–56). Hindawi Publishing Corporation.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees*. New York: Chapman & Hall, Inc.
- Candes, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2), 489–509.
- Choe, H. C., Karlsen, R. E., Meitzler, T., Gerhart, G. R., & Gorsich, D. (1996). Wavelet-based ground vehicle recognition using acoustic signals. *Proceedings of the SPIE*, 2762, 434–445.
- Coifman, R. R., Meyer, Y., & Wickerhauser, M. V. (1992). Adapted waveform analysis, wavelet-packets, and applications. In *Proceedings of ICIAM'91* (pp. 41–50). Philadelphia: SIAM Press.
- Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.
- Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4), 1289–1306.
- Donoho, D., & Tsai, Y. (2006). Extensions of compressed sensing. *Signal Processing*, 86(3), 533–548.
- Eom, K. B. (1999). Analysis of acoustic signatures from moving vehicles using time-varying autoregressive models. *Multidimensional Systems and Signal Processing*, 10, 357–378.
- Mallat, S. (1998). *A wavelet tour of signal processing*. Academic Press.
- Munich, M. E. (2004). Bayesian subspace method for acoustic signature recognition of vehicles. *Proceedings of the 12th European Signal Processing Conference, EUSIPCO*.
- Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America A*, 4(1), 519–524.
- Wickerhauser, W. V. (1994). *Adapted wavelet analysis from theory to software*. Wellesley, MA: AK Peters.
- Wu, H., Siegel, M., & Khosla, P. (1999). Vehicle sound signature recognition by frequency vector principal component analysis. *IEEE Transactions on Instrumentation and Measurement*, 48(5), 1005–1009.

Author Biographies



Amir Averbuch is professor of computer science, Tel Aviv University, Tel Aviv, Israel. He received the B.Sc and M.Sc degrees in Mathematics from the Hebrew University in Jerusalem, Israel in 1971 and 1975, respectively. He holds a Ph.D (1983) in Computer Science from Columbia University, New York. In 1977–1986 he was a Research Staff Member in the department of Computer Science, IBM Corp., T.J. Watson Research center, Yorktown Heights, New York. His research interests include applied harmonic analysis, signal/image processing, Wavelets and efficient numerical computation, fast algorithms for computational problems.



Valery A. Zheludev received his M.S. degree on Math. Physics in 1963 from St. Petersburg University, Russia. In 1968 received Ph.D. degree on Math. Physics from Steklov Math. Inst. of Acad. Sci. USSR and the prize "For the best Ph.D. thesis" at St. Petersburg University. In 1991 received Dr. Sci. degree on Comput. Math. from Siberia Branch of Acad. Sci. USSR. 1963-1965 lecturer at Pedagogical Univ., St. Petersburg, 1965-1968—Ph.D. student, St. Petersburg University, 1968–1970—Associate Professor, Kaliningrad Univ., 1970–1975—Senior Researcher, Research Inst. for Electric Measuring Devices, St.-Petersburg, 1975–1995—Associate then full Professor, St.-Petersburg Military Institute for Construction Engineering, 1995-present—Researcher then Senior Researcher, currently an Associate Professor, School of Computer Science, Tel Aviv University, Israel. Cooperated with the companies Paradigm Geophysical Ltd, Waves Audio Ltd, Bar-Kal Systems Engineering Ltd, ELTA Systems Ltd. Fields of research: Wavelet Analysis, Approximation Theory, Signal and

Image Processing, Geophysics, Pattern Recognition.



Neta Rabin received the B.Sc. degree in mathematics and computer science and the M.S. degree in computer science from Tel Aviv University, Tel Aviv, Israel, in 2002 and 2004, respectively. She is currently pursuing the Ph.D. degree at the School of Computer Science, Tel Aviv University. Her research interests are dimensionality reduction, signal and image processing.



Alon Schclar was born in Tel Aviv, Israel. He received his M.Sc. degree in computer science (summa cum laude) and his B.Sc. degree in mathematics and computer science from Tel-Aviv University, Tel Aviv, Israel, where he is currently pursuing a Ph.D. degree. In 2004, he was a fellow at the applied mathematics department at Yale University supported by the Fulbright grant for Israeli doctoral dissertation students and he was also a fellow at the Institute of Pure and Applied Mathematics at UCLA. His areas of interest include machine learning, unsupervised learning, supervised learning, dimensionality reduction, diffusion processes, signal and image processing, computer vision, computer graphics and hyper-spectral image analysis.