IEEE Instrumentation and Measurement
Technology Conference
St. Paul, Minnesota, USA, May 18-20,1998

Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis

Huadong Wu
Robotics Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh PA 15213, USA
Phone: (412) 268-2909, email: whd@cs.cmu.edu

Mel Siegel
Robotics Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh PA 15213, USA
Phone: (412) 268-8802, email: mws@cmu.edu

Pradeep Khosla
Institute for Complex Engineering Systems
Carnegie Mellon University, Pittsburgh PA 15213, USA
Phone: (412) 268-3809, email: pkk@cmu.edu

Abstract *– The sound (engine, noise, etc.) of a working vehicle provides an important clue, e.g., for surveillance mission robots, to recognize the vehicle type. In this paper, we introduce the "eigenfaces method", originally used in human face recognition, to model the sound frequency distribution features. We show that it can be a simple and reliable acoustic identification method if the training samples can be properly chosen and classified.*

We treat the frequency spectra of about 200 ms of sound (a "frame") as a vector in a high-dimensional frequency feature space. In this space, we study the vector distribution for each kind of vehicle sound produced under similar working conditions. A collection of typical sound samples is used as the training data set. The mean frequency vector of the training set is first calculated, and subtracted from each vector in the set. To capture the frequency vectors' variation within the training set, we then calculate the eigenvectors of the covariance matrix of the zero-mean-adjusted sample data set. These eigenvectors represent the principal components of the vector distribution: for each such eigenvector, its corresponding eigenvalue indicates its importance in capturing the variation distribution, with the largest eigenvalues accounting for the most variance within this data set.

Thus for each set of training data, its mean vector and its most important eigenvectors together characterize its sound signature. When a new frame (not in the training set) is tested, its spectrum vector is compared against the mean vector; the difference vector is then projected into the principal component directions, and the residual is found. The coefficients of the unknown vector, in the training set eigenvector basis subspace, identify the unknown vehicle noise in terms of the classes represented in the training set. The magnitude of the residual vector measures the extent to which the unknown vehicle sound cannot be well characterized by the vehicle sounds included in the training set.

Keywords *– sound signature, pattern recognition, frequency analysis, principal components*

I. INTRODUCTION

Almost every moving vehicle makes some kind of noise; the noise can come from the vibrations of the running engine, bumping and friction of the vehicle tires with the ground, wind effects, etc. Vehicles of the same kind and working in similar conditions ("class") will generate similar noises, or have some kind of noise signature. This noise pattern gives a clue for military reconnaissance or a surveillance mission robot to detect a vehicle and recognize its class. Our research goal is to characterize noise patterns and use them to recognize whether a new detected sound is from a vehicle of known type, and if so to classify its type.

When travelling at different speeds, under different road conditions, or with different acceleration, a vehicle emits different noise patterns. These noises can be sampled or digitized and grouped in a series of time slices (frames); then if the spectrum changes with time, it can be described in the frequency domain as the change of frequency spectrum distribution over frames.

If we consider a frame's noise frequency spectrum, with R components, as an R-dimensional vector, then each frame can be considered as a point in this R-dimensional frequency spectrum space. Noises from the same kind of vehicle and recorded under similar conditions will not be randomly distributed; if the classes are properly defined, samples from the same class should span a convex subregion, and a new sample can be classified according to its location in the frequency spectrum feature space.

To find the features in high dimensional space, we adopt and adapt the eigenfaces method used in the vision community to recognize human faces. This method is known as the Karhunen-Loeve expansion in pattern recognition, and as factor or principal-component analysis in the statistical literature.

II. SIGNAL PROCESSING

Vehicle noise is a kind of stochastic signal. A stochastic signal is defined as a stationary signal if its stochastic features are time-invariant, otherwise it is called a non-stationary signal. A vehicle that is making some noise of interest may be idling, or moving towards or away from an observing point (where the recording microphone is set); meanwhile it may be accelerating or decelerating etc. Over an extended observing time, the signal will generally not be stationary. But usually the recording microphone is fixed, and the vehicle's running conditions usually do not change very often if it is not moving; if it is moving, then a fairly short sound duration can be recorded. So vehicle sound signals can be reasonably treated as stationary, or as segments of stationary signal.

To treat the moving vehicle noise as a piece-wise stationary signal, besides the engine's running conditions, one important effect that has to be considered is the acoustic Doppler effect. The maximum Doppler effect occurs when the recording microphone is set in the vehicle path. Let $\Delta\nu$ be the Doppler frequency shift, ν be the original frequency, ΔV be vehicle travelling speed, and V be sound propagation speed; then we have $\Delta\nu/\nu = \Delta V/V$. If the vehicle is travelling at 30 m.p.h. and the speed of sound is 343.4 m/s, the maximum Doppler effect will cause about $\pm 4.2\%$ change at the frequency component ν. As the vehicle noise generally has a frequency spectrum with large low frequency components, and the recording microphone usually is set off road, the resulting Doppler shift, less than 5%, is not very conspicuous compared with the unpredictable changes in recording conditions. Experience shows that taking the sound as a stationary signal is reasonable.

Assuming each sample duration is short enough that the signal is stationary, then signal processing can be relatively simple. Below is a brief description of the process.

A. frequency analysis and spectra normalization

The recorded sound wave is digitized at a sampling rate of 22.025 kHz[1]. First, the data are normalized to 0 mean amplitude[2]. Then, the data are blocked into N frames of 4096 samples, each frame ($\overline{X}_n; n = 1, 2, \ldots, N$) sequentially with an overlap of 512 samples between adjacent frames, see figure 1 . As the engine noise can be considered as a stationary process in more than one frame (4096 sample points or 0.186 second) time interval, this 12.5% overlap is enough to smooth the result.

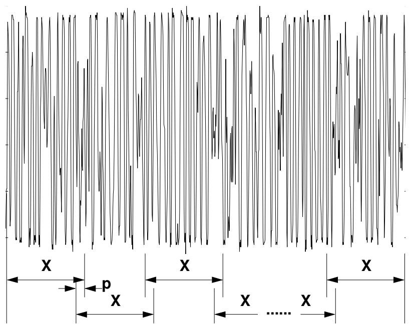

Fig. 1. Blocking sound wave samples into frames

For each complete set of samples $x_{ni}; i = 0, 1, \ldots, 4095$ in frame $\overline{X}_n$, a pre-processing smoothing filter, the Hamming window, is used to depress the Gibbs' effect in subsequent Fourier analysis:

$$w_i = 0.54 - 0.46\cos(2\pi\frac{i}{4096}), \quad i = 0, 1, \ldots, 4095 \quad (1)$$

$$x'_{ni} = x_{ni}w_i, \quad i = 0, 1, \ldots, 4095 \quad (2)$$

Next, a standard FFT algorithm is applied to each pre-processed frame. The result is a set of 4096 FFT coefficients. As the FFT phase information is not very important in sound pattern recognition, we take the spectra $SP_i, i = 0, 1, \ldots, 2047$ for subsequent analysis, i.e., we consider only the power spectrum:

$$\overline{\Phi}_n{}'' = [SP_{n0}, SP_{n1}, \ldots, SP_{n2047}]^T \quad n = 1, 2, \ldots, N \quad (3)$$

$\overline{\Phi}_n{}''$ is a vector with 2048 power spectrum components equally spaced in frequency from 5.4 Hz to 11.0125 kHz. With most vehicles, about 80% of the power spectrum is concentrated in frequencies lower than 2000 Hz, and 90% in frequencies lower than 4000 Hz. Thus to reduce computation time and memory requirement, we can take only the first 1200 components of it, that is a $\overline{\Phi}_n{}'$ is a vector with the first 1200 components of $\overline{\Phi}_n{}''$, which are the frequencies from 5.4 Hz to 6453 Hz at an increment step of 5.4 Hz.

[1] We used an ordinary tape cassette recorder to record sounds, and a Sound-Blaster card to sample the recording. The frequency response band is quite limited, but comparable to general human hearing sensitivities. 22.025 kHz is a standard SoundBlaster setting.

[2] The digitizing resolution is 8-bit. This processing removes the DC digital bias of the sound blaster card, which reports all signals in the range 0 - 255.

As the sound recording conditions are very hard to control in the field, the spectrum vectors need to be normalized before any further processing. Normalizing each frame to unit power:

$$[\phi_{n0}, \phi_{n1}, \ldots, \phi_{n1199}]^T = \frac{\overline{\Phi}_n{}'}{\sum_{i=0}^{1199} \phi_{ni}}$$

is adequate, although other schemes, e.g., normalizing it to some low stable frequency spectral component, are sometimes recommended.

B. spectrum variation adjustment

B.1 spectrum sensitivity variation over frequency

If we study the sound spectrum distribution, we can easily find that the sound spectra are generally not evenly distributed; instead, their large components heavily reside at lower end of the frequency band, and bigger variations usually accompany bigger spectrum components. Thus we need some kind of adjustment in modeling the variation of spectrum.

B.2 detection and source noise

As the frame time is short (0.186 second) at the detector end, any impulsive shaking or rubbing on the microphone causes huge variations in the frame's spectrum. At the source end, when a vehicle is moving it may experience bumps that also causes big changes in the frame's spectrum. These problems occur very often, but are not easy to pick out automatically.

Figure 2 illustrates the means and standard deviations of the frequency spectrum distribution of 2 noise samples recorded under almost the same working conditions: the microphone was at the same location, and the car was moving at about 30 m.p.h. over more-or-less the same path. It can be seen that the spectrum distributions can be quite different.

B.3 spectrum adjustment

These observations suggest that to make the analysis robust we should avoid letting small parts of spectrum variations dominate the analysis result; instead we should consider the spectrum distribution as a whole. A simple form of transformation can achieve this effect:

$$\phi_{ni} = C_2 \log_{10}(C_1 \phi_{ni}{}' + 1.0) \quad n = 1, 2, \ldots, N \tag{4}$$

$$\overline{\Phi}_n = [\phi_{n0}, \phi_{n1}, \ldots, \phi_{n1199}]^T \quad n = 1, 2, \ldots, N \tag{5}$$

The constant factors C_1 and C_2 are determined by trial-and-error experiments. For the currently available data,

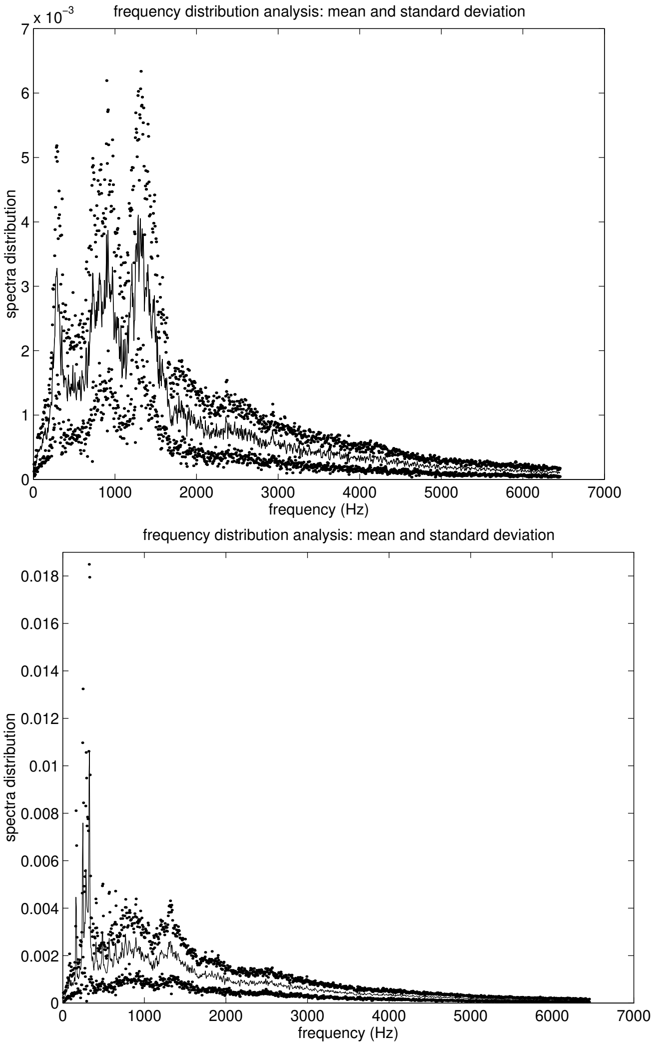

Fig. 2. spectra may vary considerably even under similar working conditions

$C_1 = 10000$ and $C_2 = 100$ give good feature abstraction, i.e., a small variation in the eigenvalues of the training set covariance matrix (described later).

III. VEHICLE NOISE PATTERN RECOGNITION

The scheme adopted here for recognition is based on an information theory approach, seeking to encode the most relevant information in a group of training samples which best distinguish them from one another. The approach transforms the noise frequency distribution variations into a small set of structures, i.e., the principal components of the initial training set of sampled noise signals.

Recognition is performed by projecting a new sample (with its mean adjusted) into the subspace spanned by the principal component structures, then by classifying the new sample as a member of the known class if its position is near the locus of that training sample set.

A. Training Processing for Pattern Feature Abstraction

Suppose we have the training set of adjusted spectrum samples $\overline{\Phi}_1, \overline{\Phi}_2, \ldots, \overline{\Phi}_N$ of the same class, i.e., from the same kind of vehicle, recorded under similar conditions. The average adjusted sound spectrum distribution of this set is defined by:

$$\overline{\Psi} = \frac{1}{N} \sum_{n=1}^{n} \overline{\Phi}_n$$

Each sample differs from the average by a variance vector $\overline{\Phi}_n - \overline{\Psi}$. This vector variance is then subject to principal component analysis, which seeks a set of M orthonormal vectors $\overline{\Theta}_k$ and their associated eigenvalues λ_k which best describe the distribution of the data. The vectors $\overline{\Theta}_k$ and scalars λ_k are the eigenvectors and eigenvalues, respectively, of the covariance matrix:

$$\frac{1}{N} \sum_{n=1}^{N} (\overline{\Phi}_n - \overline{\Psi})(\overline{\Phi}_n - \overline{\Psi})^T$$

The covariance matrix of the training set with N samples can maximally have N (in the case that $N \leq 1200$, otherwise 1200) non-trivial eigenvalues. We take the M eigenvectors $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_M$, which correspond to the M largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_M$. (It is convenient if these are appropriately arranged such that: $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_M$).

The average adjusted sound spectrum $\overline{\Psi}$ and the key eigenvectors $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_M$ of the covariance matrix together represent the main features of this vehicle sound signature. M is chosen heuristically through experiments, such that the first M largest eigenvalues are conspicuously greater than the rest of the others. Figure 3 is a typical example of an eigenvalue distribution.

B. Classification by Using Abstracted Features

Once $\overline{\Psi}$ and $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_M$ are created, a new sample can be classified by calculating how far-away the new adjusted spectrum vectors $\overline{\Gamma}_1, \overline{\Gamma}_2, \ldots, \overline{\Gamma}_P$ are from the $\overline{\Psi}$ and $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_M$ spanned subregion.

First, $\overline{\Gamma}_n, n = 1, 2, \ldots, P$ is mean-adjusted and projected onto the M orthonormal eigenvector directions:

$$\omega_{nk} = (\overline{\Gamma}_n - \overline{\Psi})^T \cdot \overline{\Theta_k} \quad k = 1, 2, \ldots, M \tag{6}$$

Then the mean and projected components are subtracted from the adjusted spectrum $\overline{\Gamma}_n - \overline{\Psi}$. The remainder is:

$$\overline{\varepsilon}_n = \overline{\Gamma}_n - \overline{\Psi} - \sum_{k=1}^{M} \omega_{nk} \overline{\Theta}_k \tag{7}$$

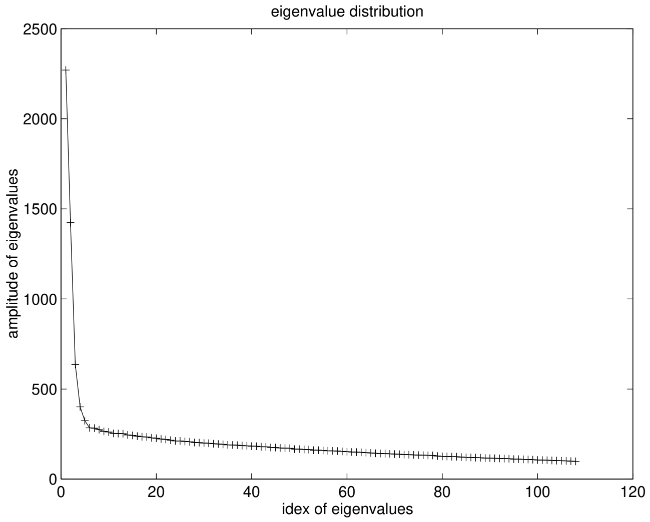

Fig. 3. typical eigenvalue distribution

The closer is the adjusted spectrum vector $\overline{\Gamma}_n$ to the feature spanned subregion, the smaller the residual components will be. So the magnitude of $\overline{\varepsilon}_n$ can be interpreted as a measurement of likelihood that $\overline{\Gamma}_n$ belongs to the class. Some threshold ε_θ can be set so that if

$$||\overline{\varepsilon}_n|| \leq \varepsilon_\theta$$

then we classify $\overline{\Gamma}_n$ as a member of the training set class, otherwise we conclude it not belongs to the class.

ε_θ is chosen by the following procedures. From the training set of adjusted spectrum vector samples, randomly choose $\overline{\tau}_1, \overline{\tau}_2, \ldots, \overline{\tau}_{N'}$. These samples are not used in the training process; instead their distances from the training set spanned subregion are measured by the residual component calculation as shown above. From their magnitude distribution ε_θ can be decided statistically.

In Figure 4 the first $\sim$30 residual magnitude-points are from the same class of cars (index 15 to 28 are from $\overline{\tau}_1, \overline{\tau}_2, \ldots, \overline{\tau}_{N'}$), the rest are from an another class a building air-conditioner.

C. Implementation

Usually for a car passing by, there can be more than 4 or 5 seconds of sustained signal available. We use a frame of about 0.2-second for each spectrum analysis, so there can be at least several dozen samples available for classification. Thus a statistical method can be used to improve the system dependability.

C.1 training example selection

An artifact of the training scheme is that to guarantee that the training group will span a convex region in fea-

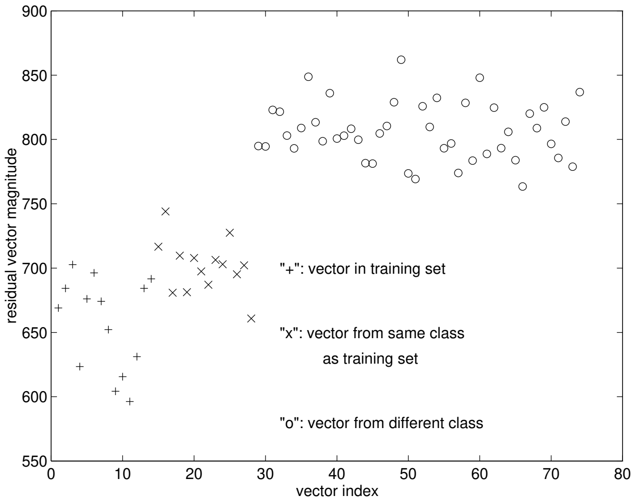

Fig. 4. typical residual distribution

ture space, we need, at the beginning of the training process, to present *only* examples that are solidly members ("core members") of the class being built. The core learning examples are those recorded under the *typical* conditions. For example, we choose sedan type cars passing the same section of road at about the same speed on sunny days (dry road surface) etc.

When new data are added to the training set, it is very important that only two sets with similar spectrum shape are merged. Otherwise the new data might smear out the features of both original data and the new data itself.

C.2 building hierarchical feature pattern

To relax recording condition constraints or to extend a known class's application range, we would hope that several groups of classes could be further generalized to form a broader class. It is indeed possible to build a hierarchical classification system structure, but only with lots of trial-and-error experiments.

For example, for sound signature extraction, the change of working and recording conditions may have greater effects than car type change. Thus it is possible that some sound signatures of different cars (travelling within certain ranges of speeds under the same road conditions) can be merged together to form a new broader class with new parameters $\overline{\Psi}$ and $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_M$; but, the sound signatures of the same kind of car can not be merged due to the variations of the weather condition (wet/dry road, wind effects, etc). The main criterion is the Euclidean distance between the means of the adjusted spectra: only two groups with small Euclidean distance between them should be merged.

Once this hierarchical structure is built, classification can be more reliable, as a new sample can checked against different range of classes.

D. Examples of Discriminating Cars from Other Vehicles

In one session of our experiments, the microphone is set to a fixed place to record all the passing vehicles' noise. Of all the recorded traffic noise data, those of sedan cars passing by at speed range 20 - 30 m.p.h. happened most often. So we choose these most typical examples to build this sound signature class. By carefully following the scheme described in the above section, we construct a model characterized by a mean spectrum vector and the six largest eigenvectors.

With this model built, we test several other types of typical vehicles. Figure 5 shows the results of a truck and a motor cycle noise. In the figure, the plus sign "+" indicates residuals from vectors in the car noise training set, the cross sign "×" indicates the residuals from vectors randomly selected from the sample class, and the small circle sign "∘" indicate those from other classes — noise of a heavy truck and a motor cycle respectively.

From the figure, it is clear that this method successfully captures the features of this sound class signature. And it is not surprised to notice that the motor cycle noise is much more easily distinguished, as it is also more significantly different in our hearing experience.

IV. RESULTS AND FUTURE RESEARCH

Under stable recording conditions, i.e., when the microphone fixed in the same place to record all samples, sound signatures of the same class can be extracted fairly reliably if we carefully follow the class feature building scheme discussed above (in Part C., section III). The above examples show a quite significant residual difference for the typical sound samples that do not belong to the known class, thus indicating this method's discrimination abilities.

With more data, we would expect the distribution difference between the training set and the test set would diminish, and thus the feature extraction to be more accurate. With more data, in Figure 4 and Figure 5, the "+" and "×" would have the same residual distribution, and it would be smaller in magnitude thus implying stronger discrimination abilities. With more data we could also have a finer discrimination between sound classes, thus means more reliably identify sounds.

The more difficult future work is to generalize our results, as to date they are more sensitive to recording conditions than we think is fundamentally necessary. We are now working towards standardizing the recording conditions and trying better equipment such as digital microphones

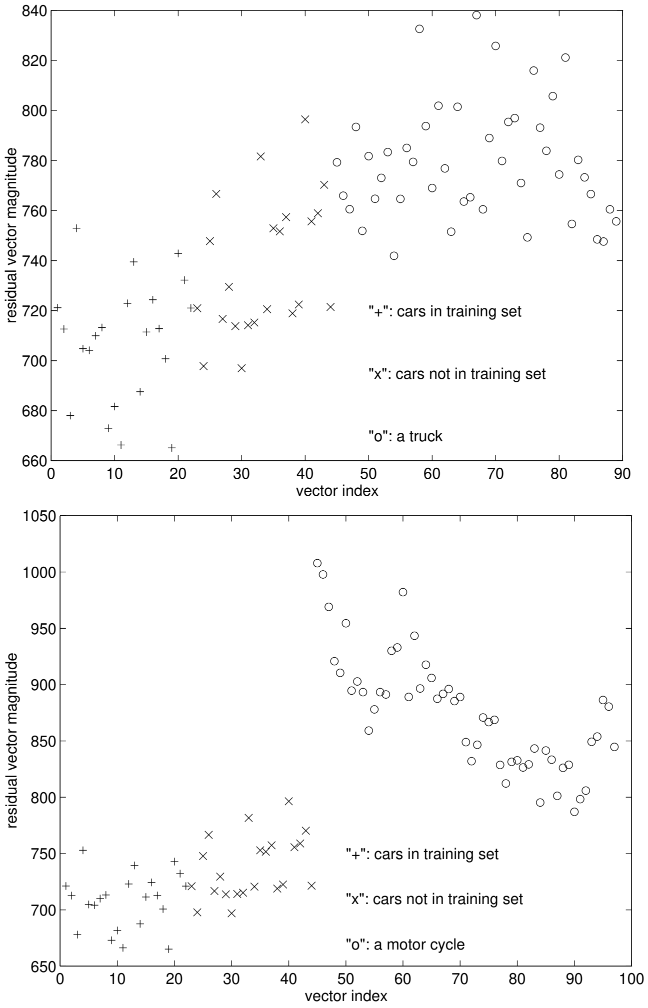

Fig. 5. classification of a heavy truck and a motor cycles from sedan car class

and recorders with higher performance. These should permit us to build a comprehensive sound signature library, and thus overcome or bypass the recording condition sensitivity problem.

The strength of using adjusted frequency spectrum principal component analysis is that a sound feature is not characterized by just a few specific frequency components; rather the whole spectrum is considered. The key requirement is to build up a properly structured, correctly classified, well-featured sound library. As this would probably be too tedious do manually for a general vehicle identification system, computer-aided supervised learning as well as feasible approaches for unsupervised learning algorithms are both necessary subjects for future research.

References

[1] Face Recognition Using Eigenfaces, Matthew A. Turk and Alex P. Pentland, 1991 IEEE

[2] Low-Dimensional Procedure for the Characterization of Human Faces, L. Sirovich and M. Kirby, Vol.4, No.3/March 1987 Journal of Optical Society of America Society (A).

[3] Human Face Recognition and Face Image Set's Topology, M. Bichsel and A. P. Pent land, CVGIP: Image Understanding, Vol. 59, No. 2, March 1994

[4] Text-Independent Speaker Identification, Herbert Gish and Michael Schmidt, IEEE Signal Processing Magazine, October 1994

[5] Patter Recognition, Vijaya Kumar, 1996