# Vehicle Sound Signature Recognition by Frequency Vector Principal Component Analysis

Huadong Wu, Mel Siegel, and Pradeep Khosla, *Fellow, IEEE*

***Abstract*—The sound of a working vehicle provides an important clue to 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 categorized. We treat the frequency spectrum in a 200 ms time interval (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 vector and the most important principal component eigenvectors of the covariance matrix of the zero-mean-adjusted samples together characterize its sound signature. When a new zero-mean-adjusted sample is projected into the principal component eigenvector directions, a small residual vector indicates that the unknown vehicle sound can be well characterized in terms of the training data set.**

***Index Terms*—Acoustic identification, frequency analysis, pattern recognition, principal components, sound signature, vehicle sounds.**

## 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.

Manuscript received November 18. This work was supported under DARPA contract F04701-97-C-0022.

H. Wu and M. Siegel are with the Robotics Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213 USA (e-mail: whd@cs.cmu.edu).

P. Khosla is with the Institute for Complex Engineering Systems, Carnegie Mellon University, Pittsburgh PA 15213 USA (e-mail: pkk@cmu.edu).

Publisher Item Identifier S 0018-9456(99)06680-2.

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 nonstationary signal. A vehicle that is making some noise of interest may be idling, or moving toward 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 a vehicle sound signal can be reasonably treated as stationary, or as segments of stationary signal.

Besides the engine's running conditions, another important effect that has to be considered, to treat the moving vehicle noise as a piecewise stationary signal, 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, $\nu$ be the original frequency, $\Delta V$ 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 50 km/h ($\sim$30 mi/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 $\nu$. 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.

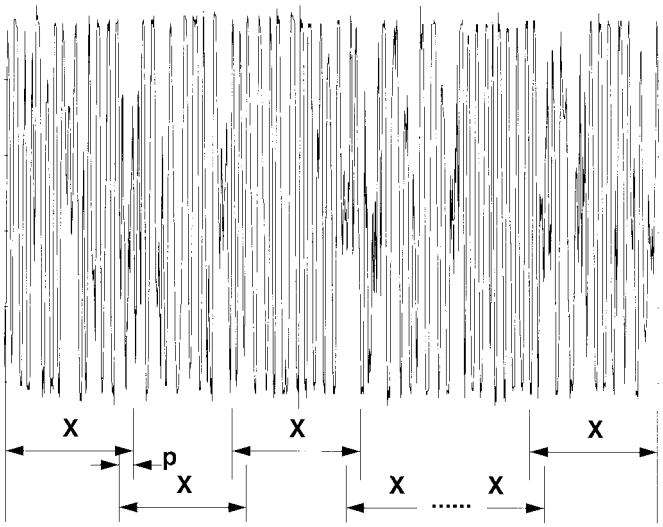


Fig. 1. Blocking sound wave samples into frames.

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 zero-mean amplitude.[2] Then, the data are blocked into $N$ frames of 4096 samples, each frame $(\bar{X}_n; n = 1, 2, \ldots, N)$ sequentially with an overlap of 512 samples between adjacent frames, see Fig. 1. As the engine noise can be considered as a stationary process in more than one frame (4096 sample points or 0.186 s) time interval, this 12.5% overlap is enough to smooth the result.

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

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

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

Next, a standard FFT algorithm is applied to each preprocessed 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 $\mathrm{SP}_i$, $i = 0, 1, \ldots, 2047$ for subsequent analysis, i.e., we consider only the power spectrum

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

where $\bar{\Phi}''_n$ 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. That is, $\bar{\Phi}'_n$ is a vector with the first 1200 components of $\bar{\Phi}''_n$, which are the frequencies from 5.4 Hz to 6453 Hz at an increment step of 5.4 Hz.

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

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

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

### B. Spectrum Variation Adjustment

*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.

*2) Detection and Source Noise:* As the frame time is short (0.186 s) 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.

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

*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 the following effect:

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

$$\bar{\Phi}_n = [\phi_{n0}, \phi_{n1}, \ldots, \phi_{n1199}]^T \quad n = 1, 2, \ldots, N. \quad (5)$$

The constant factors $C_1$ and $C_2$ are determined by trial-and-error experiments. For the currently available data, $C_1 = 10\,000$ 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

[1] We used an ordinary tape cassette recorder to record sounds, and a SoundBlaster 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.

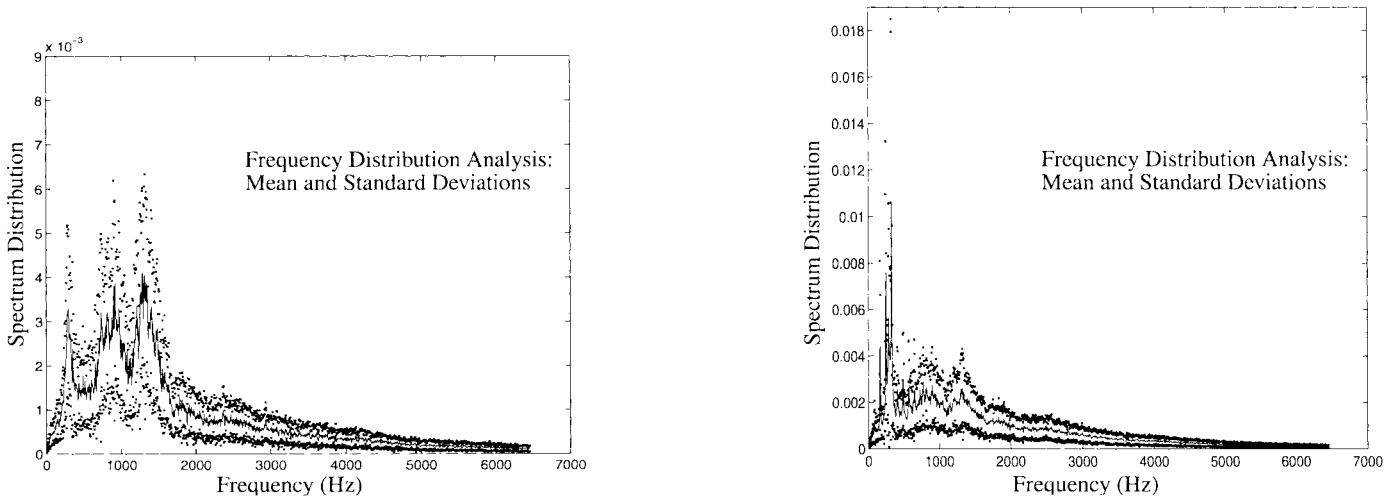


Fig. 2. Spectra may vary considerably even under similar working conditions.

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 $\bar{\Phi}_1, \bar{\Phi}_2, \ldots, \bar{\Phi}_N$ 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

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

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

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

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

The average adjusted sound spectrum $\bar{\Psi}$ and the key eigenvectors $\bar{\Theta}_1, \bar{\Theta}_2, \ldots, \bar{\Theta}_M$ 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. Fig. 3 is a typical example of an eigenvalue distribution.

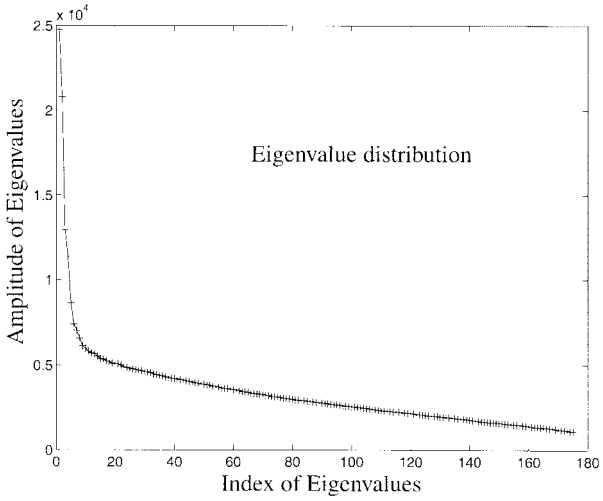


Fig. 3. Typical eigenvalue distribution.

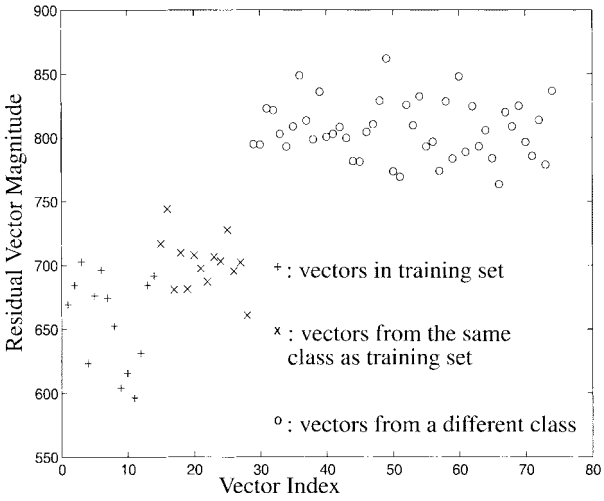


Fig. 4. Typical residual distribution.

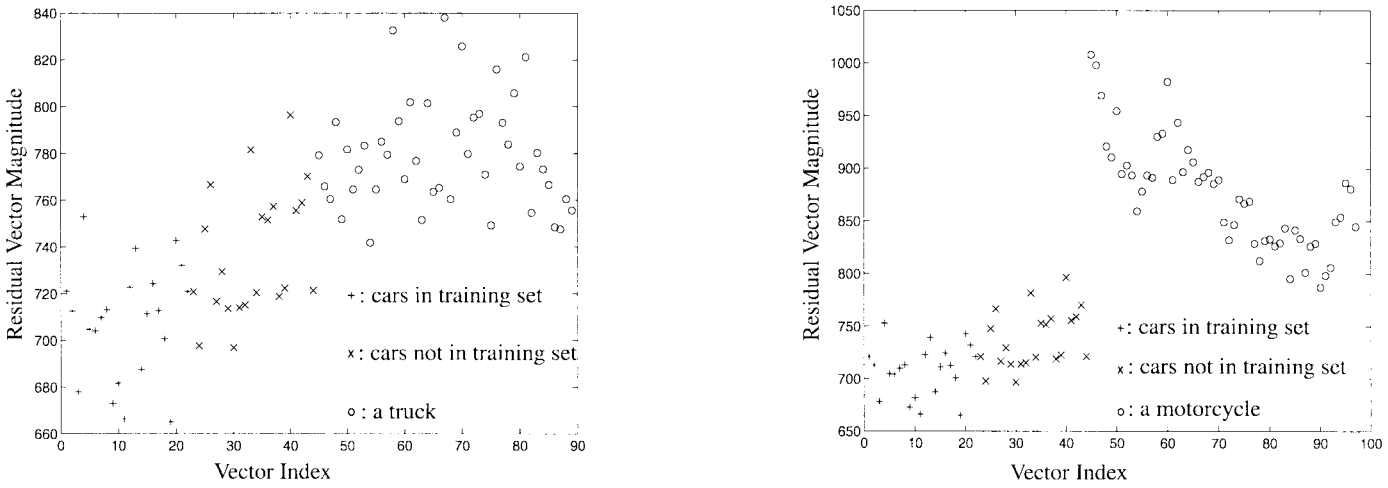


Fig. 5. Classification of a heavy truck and a motor cycles from sedan car class.

### B. Classification by Using Abstracted Features

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

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

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

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

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

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

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

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

$\varepsilon_\theta$ is chosen by the following procedures. From the training set of adjusted spectrum vector samples, randomly choose $\bar{\tau}_1, \bar{\tau}_2, \ldots, \bar{\tau}_{N'}$. 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 $\varepsilon_\theta$ can be decided statistically.

In Fig. 4 the first $\sim$30 residual magnitude-points are from the same class of cars (index 15 to 28 are from $\bar{\tau}_1, \bar{\tau}_2, \ldots, \bar{\tau}_{N'}$), 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 s sustained signal available. We use a frame of about 0.2 s 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.

*1) Training Example Selection:* An artifact of the training scheme is that to guarantee that the training group will span a convex region in feature 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.

*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 $\bar{\Psi}$ and $\bar{\Theta}_1, \bar{\Theta}_2, \ldots, \bar{\Theta}_M$; 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 be checked against different ranges 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 30–50 km/h (about 20–30 mi/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. Fig. 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., with 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 in Section III-C. 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 Figs. 4 and 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, so 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 toward standardizing the recording conditions and trying better equipment such as digital microphones 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] M. A. Turk and A. P. Pentland, *Face Recognition Using Eigenfaces*. New York: IEEE Press, 1991.
[2] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer. Soc. A*, vol. 4, Mar. 1987.
[3] M. Bichsel and A. P. Pentland, "Human face recognition and face image set's topology," *CVGIP: Image Understanding*, vol. 59, Mar. 1994.
[4] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, Oct. 1994.
[5] V. Kumar, "Pattern recognition," unpublished.

**Huadong Wu** received the B.S. and M.S. degrees in precision instrumentation in 1984 and 1987, respectively, from Shanghai Jiaotong University, Shanghai, China. He is pursuing the Ph.D. degree at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

From 1987 to 1995, he was an engineer with the Robot and Automation Research Institute, Beijing, China, primarily in robot mechanical designing and the control of robot manufacturing systems.

**Mel Siegel** received the B.A. degree from Cornell University, Ithaca, NY, and the M.S. and Ph.D. degrees from the University of Colorado, Boulder, all in physics.

He is a Senior Research Scientist at the Robotics Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. He is also Director of the Sensors, Measurement and Control Laboratory. His background is in physical and measurement sciences, analytical instrument development, and the application of computer science to measurement, diagnosis, and control; he is also interested in 3-D stereoscopic video and computer graphics display systems. The dominant themes in his recent research have been "difficult measurements in difficult environments," exemplified by his use of mobile robots to inspect aging aircraft, and high definition technology as enabling means for a new generation of 3-D stereoscopic technologies and applications.

**Pradeep Khosla** (S'83–M'83–SM'91–F'95) received the B.Tech.(Hons.) degree from the Indian Institute of Technology, Kharagpur, India, and the M.S. and Ph.D. degrees from Carnegie Mellon University, Pittsburgh, PA, in 1984 and 1986, respectively.

He joined Carnegie Mellon University in 1986 and is currently Professor of electrical and computer engineering and robotics and Founding Director of the Institute for Complex Engineered Systems.

Dr. Khosla was the Program Vice-Chairman for the 1989 IEEE International Conference on Systems Engineering, General Chairman for the 1990 IEEE International Conference on Systems Engineering, Program Vice Chairman of the 1993 International Conference on Robotics and Automation, General Co-Chairman of the 1995 Intelligent Robotics Systems (IROS) conference, and Program Vice-Chair for the 1997 IEEE Robotics and Automation Conference. He has served as member of the AdCom of the IEEE Systems, Man and Cybernetics Society, Associate Editor of the IEEE Transactions on Robotics and Automation, and Chairman of the Education Committee of the IEEE Robotics and Automation Society.