Acoustic Signal based Traffic Density State Estimation using Adaptive Neuro-Fuzzy Classifier

PRASHANT BORKAR[#1], L. G. MALIK[#2], M. V. SARODE[#3]
DEPARTMENT OF CSE
[#1, #2] G.H. RAISONI COLLEGE OF ENGINEERING, NAGPUR, INDIA
[#3] BABASAHEB AMBEDKAR COLLEGE OF ENGINEERING AND RESEARCH, NAGPUR, INDIA

prashant.borkar@raisoni.net, latesh.malik@raisoni.net, milind09111970@gmail.com

Abstract— Traffic monitoring and parameters estimation from urban to battlefield environment traffic is fast-emerging field based on acoustic signals. This paper considers the problem of vehicular traffic density state estimation, based on the information present in cumulative acoustic signal acquired from a roadside-installed single microphone. The occurrence and mixture weightings of traffic noise signals (Tyre, Engine, Air Turbulence, Exhaust, and Honks etc) are determined by the prevalent traffic density conditions on the road segment. In this work, we extract the short-term spectral envelope features of the cumulative acoustic signals using MFCC (Mel-Frequency Cepstral Coefficients). The (Scaled Conjugate Gradient) SCG algorithm, which is a supervised learning algorithm for network-based methods, is used to computes the second-order information from the two first-order gradients of the parameters by using all the training datasets. Adaptive Neuro-Fuzzy classifier is used to model the traffic density state as Low (40 Km/h and above), Medium (20-40 Km/h), and Heavy (0-20 Km/h). For the developing geographies where the traffic is non-lane driven and chaotic, other techniques (magnetic loop detectors) are inapplicable. Adaptive Neuro-Fuzzy classifier is used to classify the acoustic signal segments spanning duration of 20–40 s, which results in a classification accuracy of 93.33% for 13-D MFCC coefficients and around 96% when entire features were considered, 77.78% for first order derivatives and ~75% for second order derivatives of cepstral coefficients.

Keywords: Acoustic signal, Noise, Traffic, Density, Neuro-Fuzzy.

1. Introduction

As the number of vehicle in urban areas is ever increasing, it has been a major concern of city authorities to facilitate effective control of traffic flows in urban areas [1]. Especially in rush hours, even a poor control at traffic signals may result in a long time traffic jam causing a chain of delays in traffic flows and also CO2 emission [2]. Density of traffic on roads and highways has been increasing constantly in recent years due to motorization, urbanization, and population growth. Intelligent traffic management systems are needed to avoid traffic congestions or accidents and to ensure safety of road users.

Traffic in developed countries is characterized by lane driven. Use of magnetic loop detectors, video cameras, and speed guns proved to be efficient approach for traffic monitoring and parameter extraction but the installation, operational and maintenance cost of these sensors significantly adds to the high operational expense of these devices during their life cycles. Therefore researchers have been developing several numbers of sensors, which have a number of significant advantages and disadvantages relative to each other. Nonintrusive traffic-monitoring technologies based on ultrasound, radar (Radio, Laser, and Photo), video and audio signals. All above present different characteristics in terms of robustness to changes in environmental conditions; manufacture, installation, and repair costs; safety regulation compliance, and so forth [3].

Traffic surveillance systems based on video cameras cover a broad range of different tasks, such as vehicle count, lane occupancy, speed measurements and classification, but they also detect critical events as fire and smoke, traffic jams or lost cargo. The problem of traffic monitoring and parameter estimation is most commonly solved by deploying inductive loops. These loops are very intrusive to the road pavement and, therefore cost associated with these is very high. Most video analytics systems on highways focus on counting and classification [4], [5], [6], [7], [8]. Key requirement for any video based traffic monitoring system is ability to handle varied lightning condition

and occlusions in heterogeneous network. References [30, 40-42] describe some latest technologies which are robust traffic monitoring using video, ranging from vehicle tracking to vehicle occlusion handling.

For detecting vehicles in urban traffic scenes by means of rule-based reasoning on visual data, Cucchiara et al. [43] proposed an approach. In [44], Kamijo et al. proposed a hidden Markov model (HMM)-based computer-vision technique to detect accidents and other events such as reckless driving at road intersections. However, the problem of average-speed/speed-range estimation is not directly address. Coifman et al. proposed an extensive feature-based computer-vision technique for vehicle tracking. They use the "corner" features of the vehicles, which are being driven in the lanes, to track them and then estimate traffic parameters such as average speed and volume. They obtained impressive results on free-way traffic, where more that 80% vehicles were traveling within the speed range of 50–70 mi/h (80–110 km/h) [45]. These speeds leads to good tracking as the vehicles are not linked to each other. However, it is not clear if such a tracking technique could still work in the chaotic and nonlane-driven city traffic conditions with the extremely varied speed ranges of 0–20, 20–40 km/h, and more than 40 km/h.

Such traffic conditions are very common in cities of developing geographies (India and South Asia) and are the focus of this paper. Using general purpose surveillance cameras for traffic analysis is demanding job. The quality of surveillance data is generally poor, and the range of operational conditions (e.g., night time, inclement, and changeable weather) requires robust techniques. The use of *road side acoustic signal* seems to be good approach for traffic monitoring and parameter estimation purpose having very low installation, operation and maintenance cost; low-power requirement; operate in day and night condition.

Conventional pattern classification involves clustering training samples and associating clusters to given categories with limitations of lacking of an effective way of defining the boundaries among clusters. On the contrary, fuzzy classification assumes the boundary between two neighboring classes as a continuous, overlapping area within which an object has partial membership in each class [9]. In brief, we use fuzzy IF-THEN rules to describe a classifier.

Assume that K patterns, $x_p = (x_{p1,\ldots\ldots}x_{p2})$, p= 1, .. K are given from two classes, where x_p is an n-dimensional crisp vector. Typical fuzzy classification rules for n = 2 are like

If x_{p1} is small and x_{p2} is very large then
$x_p = (x_{p1}, x_{p2})$ belongs to C1
If x_{p1} is large and x_{p2} is very small then
$x_p = (x_{p1}, x_{p2})$ belongs to C2

Where x_{p1} and x_{p2} are the features of pattern (or object) p, small and very large are linguistic terms characterized by appropriate membership functions. The firing strength or the degree of appropriateness of this rule with respect to a given object is the degree of belonging of this object to the class C.

Most of the classification problems consist of medium and large-scale datasets, example: genetic research, character or face recognition. For this different methods, such as neural networks (NNs), support vector machines, and Bayes classifier, have been implemented to solve these problems. The network-based methods can be trained with gradient based methods, and the calculations of new points of the network parameters generally depend on the size of the datasets. One of the network-based classifiers is the Neuro-Fuzzy Classifier (NFC), which combines the powerful description of fuzzy classification techniques with the learning capabilities of NNs.

The Scaled Conjugate Gradient (SCG) algorithm is based on the second-order gradient supervised learning procedure [10]. The SCG executes a trust region step instead of the line search step to scale the step size. The line search approach requires more parameters to determine the step size, which results in increasing training time for any learning method. In a trust region method, the distance for which the model function will be trusted is updated at each step. The trust region methods are more robust than line-search methods. The disadvantage associated with line-search method is eliminated in the SCG by using the trust region method [10].

We start with a characterization of the road side cumulative acoustic signal which comprising several noise signals (tire noise, engine noise, air turbulence noise, and honks), the mixture weightings in the cumulative signal varies, depending on the traffic density conditions [11]. For low traffic conditions, vehicles tend to move with medium to high speeds, and hence, their cumulative acoustic signal is dominated by tire noise and air turbulence noise [11], [12]. On the other hand, for a heavily congested traffic, the acoustic signal is dominated by engine-idling noise and the honks. Therefore, in this work, we extract the spectral features of the roadside acoustic signal using Mel-Frequency Cepstral Coefficients (MFCC), and then Adaptive Neuro-Fuzzy Classifier is used to determine the traffic density state (low, Medium and

Heavy). This results in 93.33% accuracy when 20–30 s of audio signal evidence is presented.

We begin with description of the various noise signals in the cumulative acoustic signal in Section 2. Overview of past work based on acoustic signal for traffic monitoring is provided in Section 3, followed by feature extraction using Mel-Frequency Cepstral Coefficients in 4. Finally, the experimental setup and the classification results by ANFC are provided in Section 5, and the conclusion is summarized in Section 6.

2. Vehicular Acoustic Signal

A vehicular acoustic signal is mixture of various noise signals such as tyre noise, engine idling noise, noise due to exhaust, engine block noise, noise due to aerodynamic effects, noise due to mechanical effects (e.g., axle rotation, brake, and suspension), air-turbulence noise and the honks. The mixture weighting of spectral components at any location is depends upon the traffic density condition and vehicle speed. In former case if we consider traffic density as freely flowing then acoustic signal is mainly due to tyre noise and air turbulence noise. For medium flow traffic acoustic signal is mainly due to wide band drive by noise, some honks. For heavy traffic condition the acoustic signal is mainly due to engine idling noise and several honks. A typical vehicle produces various noise depends on its velocity, load and mechanical condition. In general, approximation can be done as vehicular acoustic signal is categorized as,

2.1 Tyre noise

Tyre noise refers to noise produced by rolling tyre as an interaction of rolling tyre with road surface. The tyre noise is also considered as main source of vehicle's total noise at a speed higher than 50 kph [12], [13]. Tyre noise has two components: air noise and vibrational noise [13], [14]. Air noise dominant in the frequency ranges between 1 KHz to 3 KHz. On the other hand vibrational noise is dominant in the frequency range 100 Hz to 1000 Hz. Effect is generated by road and tyre, which forms a geometrical structure that amplifies the noise (amplification results in tyre noise component in the frequency range 600 Hz to 2000 Hz), produced due to tyre-road interaction [14], [15], [16]. The directivity of horn depends upon tyre geometry, tyre thread geometry, weight and torque of tyre. The total tyre noise power along with horn effect lies in the frequency range 700–1300 Hz.

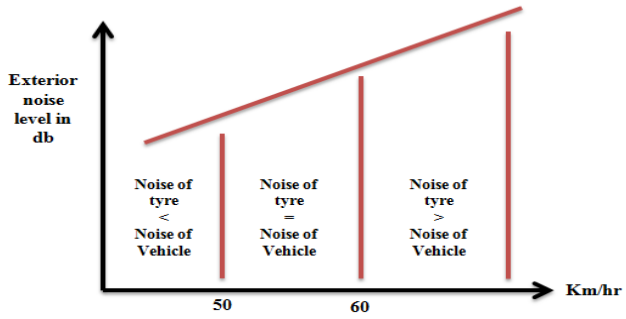
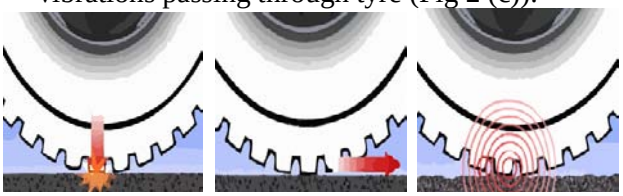


Fig. 1. Noise of the tyre Vs Noise of the vehicle.

The tyre noise is caused by three different factors:

- Tyre hitting ground (Fig 2 (a)).
- Vibration of air through tread pattern (Fig 2 (b)).
- vibrations passing through tyre (Fig 2 (c)).



(a) (b) (c)

Fig. 2. (a) Tyre hitting the ground, (b) Vibration of the air through the tread pattern, (c) Vibrations passing through the tyre

2.2 Engine noise

Engine noise is produced due to internal combustion of engine. Engine noise contains a deterministic harmonic train and stochastic component due to air intake [11]. The fuel combustion in engine cylinder leads to deterministic harmonic train where lowest harmonic tone refers to cylinder fire rate. On the other hand stochastic component is largely due to the turbulent air flow in the air intake, the engine cooling systems, and the alternator fans. The engine noise varies with speed and the acceleration of vehicle [11], [17]. A stationary vehicle produces distinct engine idling noise whereas moving vehicle produces different engine noise in correspondence with cylinder fire rate. In the recent years, manufacturers designs quieter engine to suppress the noise level. So engine noise might be strong on front side of car compared to other directions.

2.3 Exhaust noise

The exhaust noise is produced due to entire exhaust system. The system goes from the engine combustion compartment through exhaust tubes to the exhaust muffler present at the back of the vehicle generating exhaust noise. The exhaust noise is directly proportional to load of the vehicle [18]. The exhaust noise is characterised by having power spectrum around lower frequencies. Exhaust noise is

affected by turbo chargers and after cooler [18], [19].

2.4 Air Turbulence noise

Air turbulence noise is produced due to the air flow generated by the boundary layer of the vehicle. It is prominent immediately after the vehicle passes by the sensor (e.g. microphone). It produces distinctive drive-by-noise or *whoosh* sound. The Air turbulence noise depends on the aerodynamics of the vehicle, wind speed and its orientation [20], [21].

3. Acoustic Signals for Traffic Monitoring

Today's urban environment is supported by applications of computer vision techniques and pattern recognition techniques including detection of traffic violation, vehicular density estimation, vehicular speed approximation, and the identification of road users. Currently magnetic loop detector is most widely used sensor for traffic monitoring in developing countries [22]. However traffic monitoring by using these sensors still have very high installation and maintenance cost. This not only includes the direct cost of labor intensive earth work but also, perhaps more importantly, the indirect cost associated with the disruption of traffic flow. Also these techniques require traffic to be orderly flow, traffic to be lane driven and in most cases it should be homogeneous.

Referring to the developing regions such India and Asia the traffic is non lane driven and highly chaotic. Highly heterogeneous traffic is present due to many two wheelers, three wheelers, four wheelers, auto-rickshaws, multi-wheeled buses and trucks, which does not follow lane. So it is the major concern of city authority to monitor such chaotic traffic. In such environment the loop detectors and computer-vision-based tracking techniques are ineffective. The use of road side acoustic signal seems to be good alternative for traffic monitoring purpose having very low installation, operation and maintenance cost.

3.1 Vehicular Speed Estimation

Doppler frequency shift is used to provide a theoretical description of single vehicle speed. Assumption made that distance to the closest point of approach is known the solution can accommodate any line of arrival of the vehicle with respect to the microphone. The description applies only to a single vehicle's acoustic waveform and in case of several vehicles the interference of their mixed acoustic waveforms will render this solution inapplicable to their speed estimation [23], [24].

Sensing techniques based on passive sound detection are reported in [25], [26]. These techniques utilizes microphone array to detect the sound waves generated by road side vehicles and are capable of capable of monitoring traffic conditions on lane-by-lane and vehicle-by-vehicle basis in a multilane carriageway. S. Chen et Al develops multilane traffic sensing concept based passive sound which is digitized and processed by an on-site computer using a correlation based algorithm. The system having low cost, safe passive detection, immunity to adverse weather conditions, and competitive manufacturing cost. The system performs well for free flow traffic however for congested traffic performance is difficult to achieve [27].

Valcarce *et al*. exploit the differential time delays to estimate the speed. Pair of omnidirectional microphones was used and technique is based on maximum likelihood principle [3]. Lo and Ferguson develop a nonlinear least squares method for vehicle speed estimation using multiple microphones. Quasi-Newton method for computational efficiency was used. The estimated speed is obtained using generalized cross correlation method based on time-delay-of-arrival estimates [28].

Cevher *et al.* uses single acoustic sensor to estimate vehicle's speed, width and length by jointly estimating acoustic wave patterns. Wave patterns are approximated using three envelop shape components. Results obtained from experimental setup shows the vehicle speeds are estimated as (18.68, 4.14) m/s by the video camera and (18.60, 4.49) m/s by the acoustic method [29]. They also had estimated a single vehicle's speed, engine's rounds per minute (RPM), the number of cylinders, and its length and width based on its acoustical wave patterns [17]. However, their technique is applicable only when there is single vehicle travelling on the road and its vehicle type has to be recognized (such as Ford F150, Honda Accord, VW Passat etc). Therefore, it cannot be applied for traffic density state estimation where there are multiple vehicles travelling and producing a cumulative acoustic signal rather than just a single vehicle's acoustic signal.

Combination of smartphone features such as accelerometers and basic honk signal detection, followed by a simple Doppler frequency shift computation, to arrive at a vehicle's speed estimate [46]. In [47], Lee and Rakotonirainy presented an approach for detecting a crash-risk level using the computing power and the microphones of mobile devices that can be used to alert the user in advance of an approaching vehicle to avoid a crash.

3.2 Traffic Density Estimation

Urban areas are concerned with effective traffic signal control and traffic management. Time estimation for reaching from source to destination using real time traffic density information is major concern of city authorities. Referring to the developing geographical areas like Asia, the traffic is characterised be non lane-driven. In such condition traffic density estimation using magnetic loop detectors, speed guns and video monitoring seems to be best, but the installation, maintenance and operation cost associated with these approaches are very high. Use of road side acoustic signal seems to be an alternative for traffic density estimation. Jien Kato proposed method for traffic density estimation based on recognition of temporal variations that appear on the power signals in accordance with vehicle passes through reference point [30]. HMM is used for observation of local temporal variations over small periods of time, extracted by wavelet transformation. Experimental results show good accuracy for detection of passage of vehicles.

Vivek Tyagi *et al.* classify traffic density state as free flowing, Medium flow and Jammed. They consider short term spectral envelops features of cumulative acoustic signal, and then class conditional probability distribution is modelled on one of the three broad traffic density state (mentioned above). Experimental setup uses omnidirectional microphone placed at about 1.5 m height and cumulative acoustic signal is recorded at 16000 Hz sampling frequency. Bayes classifier is applied to classify traffic density state which results in ~ 95% of accuracy, which is then improved by using discriminative classifier such as RBF-SVM [48]. Compare with the existing computer vision and traffic monitoring system in [49] and [50] this technique is independent of light condition and works well for developing regions. Techniques in [46], [51] requires accurate detection of honk signal to arrive at average speed. However [48] can't provide very accurate speed estimation compared to.

3.3 Vehicular Classification

Problem of vehicular classification is example of pattern recognition theory. Acoustic signals collected by acoustic sensors are used to identify the type of moving ground vehicles. Typical classification process consists of sensing, class definition, feature extraction, classifier application and system evaluation. Based on collected acoustic data feature vectors are extracted.

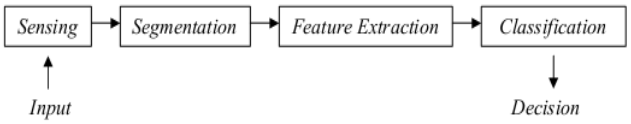


Fig. 3. Typical classification system (ref. [52])

Referring to the figure 3 Sensing unit collects raw data in order to provide sensor node the information about traffic condition. Segmentation refers to separation of single vehicle imposes major restriction on acoustic classification system because traffic recordings are consist of signals from multiple vehicles which are mutually overlap. Feature extraction refers to extracting representative set of features which are able to distinguish different classes of vehicle. Richard O. Duda writes in "Pattern Classification" [52]: "The conceptual boundary between feature extraction and classification proper is somewhat arbitrary: An ideal feature extractor would yield a representation that makes the job of the classifier trivial; conversely, an omnipotent classifier would not need the help of a sophisticated feature extractor. The distinction is forced upon us for practical rather than theoretical reasons." Classification decides which class or category a given feature vector belongs. Many classifiers do this by supervised learning, where a representative training set of feature vectors for each class is used to train or learn the classifier. Classification learning schemes usually use one of the following approaches:

- *Statistical classifiers* based on Bayes decision theory, assume an underlying probability distribution for unknown patterns, e.g. maximum likelihood estimation, maximum posterior probability estimation, Gaussian mixture models, hidden Markov models or k-nearest neighbor method.
- *Syntactic or structural classifiers* based on linear or nonlinear interrelationships of features in the feature vector lead to linear/non-linear classifier.

Acoustic feature generation are mainly based on three domains: time, frequency, and both time-frequency domain.

- *Time domain feature* generation offers very low computational demand, but features are often hampered by environmental noise or wind effects.
- *Frequency domain feature generation* consider a stationary spectrum in a given time frame. As moving vehicles are non-stationary signals, the influence of Doppler effects and signal energy changes either have to be neglected or the

investigated time frame must be chosen short enough to afford quasi stationary signal behavior.

- *Time-frequency domain feature generation* consider the non-stationary signal behavior of passing vehicles and it lead to accurate measures of signal energies in time and frequency domain simultaneously, these approaches are having a high computational complexity.

Table 1. Vehicular acoustic feature extractors and classifiers

Domain	Ref.	Feature Extractor	Classifier used	Accuracy
Time	[31]	TE, PCA	Fuzzy Logic, MLNN	73-79% 95-97.5%
	[32]	Correlation based algorithm		
Frequency	[33]	HLA	NN	Vehicle: 88% Cylinder: 95%
	[34]	HLA, DWT, STFT, PCA	k-NNS, MPP	kNN: 85% MPP: 88%
	[35]	AR mod.	MLNN	up to 84%
Time-Frequency	[36]	DWT	MPP	98.25%
	[34]	HLA, DWT, STFT, PCA	k-NNS, MPP	kNN: 85% MPP: 88%

Table 2. Acronyms from section 3 and 4

TE Time Energy Distribution	**MLNN** Multi Layer Neural Network.
PCA Principal Components Analysis	**NN** Artificial Neural Network
HLA Harmonic Line Association	**k-NNS** k – Nearest Neighbor Search
DWT Discrete Wavelet Transform	**MPP** Maximum Distance Approach
STFT Short Time Fourier Analysis	**AR mod.** Autoregressive Modeling
CWT Continuous Wavelet Transform	

3.4 Proposed Approach

Given limitations literature, we propose to use the entire cumulative roadside acoustic signal rather than just detecting the honk signal. Section 4, we will show, through the spectral analysis, that the cumulative acoustic signal has important discriminative information present in its spectro-temporal plane that allows us to directly build simple statistical classifiers to classify between three broad traffic density states that correspond to an increasing range of speeds, i.e., (0, 20) km/h, (20, 40) km/h, and above 40 km/h. We denote them as Heavy (Jammed), Medium-Flow, and Free-Flow traffic density states, respectively.

One of the main characteristic of the city traffic in the developing geographies (particularly South Asia/India) is that it usually does not move in the lanes, even if they are explicitly marked on the roads. Frequent lane changing is very common, and hence, a lane-based volume measure (number of vehicles passing a point of a lane per hour) does not seem like an appropriate measure in such conditions. The entire road width, with all the lanes combined, becomes one continuous carriage-way. Therefore, we have decided to use the measure of traffic density (Heavy, Medium-Flow, and Free-Flow corresponding to an increasing range of speeds (0, 20) km/h, (20, 40) km/h, and above 40 km/h, respectively), instead of the two distinct volume and speed measures, as has been used in some of the traffic monitoring work in the developed geographies.

4. Feature Extraction using MFCC

An omnidirectional microphone was placed on the pedestrian sidewalk at about 1 to 1.5 m height. We have collected about 3 hr of cumulative roadside acoustics data from the Chattrapati Square to T-point Nagpur, India. Samples were collected for time durations of around 30s for different traffic density state conditions (low, medium and heavy). The data were collected from a roadside installed omnidirectional microphone at 16-kHz sampling frequency. These data covered three broad traffic density classes and were collected from about five different road segments. The labelling of the data was done by a human assessment of the prevailing traffic density state. We further partitioned the data into two independent sets, i.e., one for training the parameters of the three classes (traffic density states) and another for the classification experiments based on the learned distributions as in. The training set equally covered three traffic density states (classes) and consisted of 900 sec of audio data. Similarly, the test set was of duration 900 sec and almost equally covered the three classes.

The various traffic density states induce different cumulative acoustic signals. To prove the above statement, we have examined the spectrogram of the different traffic state's cumulative acoustic signals.

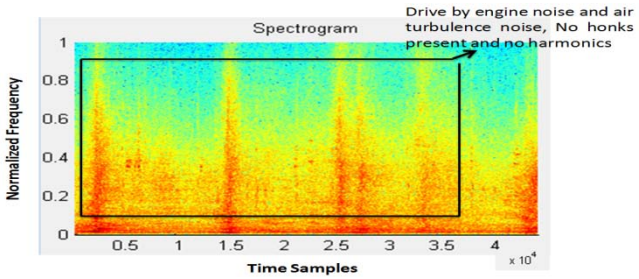


Fig. 4. Spectrogram of the low density traffic (above 40 km/h).

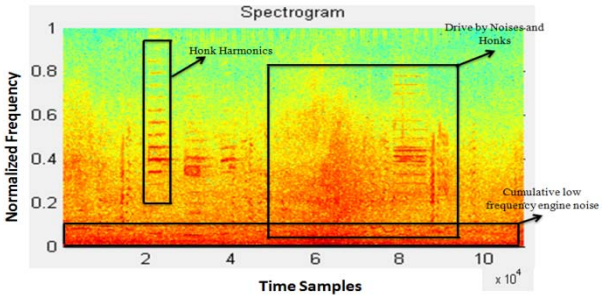


Fig. 5. Spectrogram of the Medium density traffic (20 to 40 km/h).

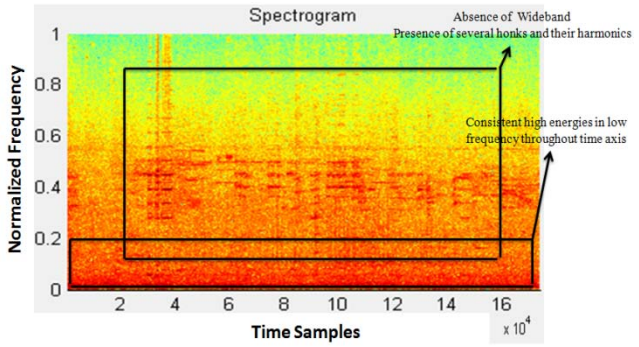


Fig. 6. Spectrogram of the Heavy density traffic (0 to 20 km/h).

- For the *low density traffic* condition in Fig. 4, we only see the wideband drive-by noise and the air turbulence noise of the vehicles. No honks or very few honks are observed for low density traffic condition.
- For the *medium density traffic* condition in Fig. 5, we can see some wideband drive-by noise, some honk signals, and some concentration of the spectral energy in the low-frequency ranges (0, 0.1) of the normalized frequency or equivalently (0, 800) Hz.
- For the *heavy density traffic* condition in Fig. 6, we notice almost no wideband drive-by engine noise or air turbulence noise and are dominated by several honk signals. We note the several harmonics of the honk signals, and they are ranging from (2, 6) kHz.

The goal of feature extraction is to give a good representation of the vocal tract from its response characteristics at any particular time. Mel-Frequency cepstral coefficients (MFCC), which are the Discrete Cosine Transform (DCT) coefficients of a Mel-filter smoothed logarithmic power spectrum. First 13–20 cepstral coefficients of a signal's short time spectrum succinctly capture the smooth spectral envelope information. We have decided to use first 13 cepstral coefficients to represent acoustic signal for corresponding traffic density state. These coefficients have been very successfully applied as the acoustic features in speech recognition, speaker recognition, and music recognition and to vast variety of problem domains. Features extraction using MFCC is as follows,

A. Pre-emphasis

Pre-emphasis phase emphasizes higher frequencies. The pre-emphasis is a process of passing the signal through a filter. It is designed to increase, within a band of frequencies, the magnitude of some frequencies (higher) with respect to the magnitude of the others frequencies (lower) in order to improve the overall SNR.

$$y[n]= x[n]-\alpha x[n-1],\ \alpha \in (0.9, 1) \quad (1)$$

Where x[n] denotes input signal, y[n] denotes output signal and the coefficient α is in between 0.9 to 1.0, α= 0.97 usually. The goal of pre-emphasis phase is to compensate high-frequency part that was suppressed during the sound collection.

B. Framing and Windowing

Typically, speech is a non-stationary signal; therefore its statistical properties are not constant across time. The acquired signal is assumed to be stationary within a short time interval. The input acoustic signal is segmented in frames of 20~40 ms with overlap (optional) of 1/3~1/2 of the frame size. In order to keep the continuity of the first and the last points in the frame, typically each frame has to be multiplied with a hamming window. Its equation is as follows,

$$W[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N}, 0 \le n \le N \\ 0, otherwise \end{cases} \quad (2)$$

Where N is frame size

$$Y[n]= X[n] * W[n] \quad (3)$$

Where Y[n] = Output signal

X[n] = Input signal
W[n] = Hamming Window

Due to the physical constraints, the traffic density state could change from one to another (low to medium flow to heavy) over at least 5–30 min duration. Therefore, we decided to use relatively longer primary analysis windows of the typical size 500 ms and shift size of 100 ms to obtain the spectral envelope.

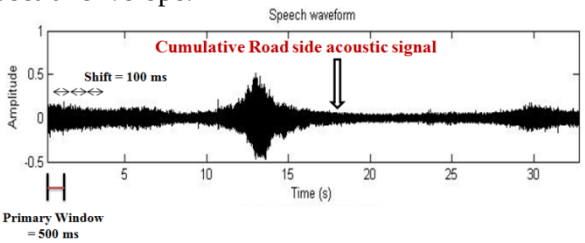


Fig. 7. Primary windows of size=500 ms and shifted by 100 ms to obtain a sequence of MFCC feature vectors.

C. DFT

Commonly, Fast Fourier Transform (FFT) is used to compute the DFT. It converts each frame of N samples from time domain into frequency domain. The computation of the FFT-based spectrum as follow,

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}}, 0 \leq k \leq N \quad (4)$$

Where N is the frame size in samples, x[n] is the input acoustic signal, and. X[k] is the corresponding FFT-based spectrum.

D. Triangular bandpass filtering

The frequencies range in FFT spectrum is wide and acoustic signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and is equal to unity at the Centre frequency and decrease linearly to zero. We then multiply the absolute magnitude of the DFT samples by the triangular frequency responses of the 24 Mel-filters that have logarithmically increasing bandwidth and cover a frequency range of 0–8 kHz in our experiments. Each filter output is sum of its filtered spectral components. To compute the Mel for given frequency f in HZ, equation is as follows:

$$F(Mel) = 2595 * \log 10 [1+f/700] \quad (5)$$

The *i*th Mel-filter bank energy $(M_{FB}(i))$ is obtained as

$$(M_{FB}(i)) = (Mel_i(k)) * |X(k)|^2, k \in (0, N/2) \quad (6)$$

Where $(Mel_i(k))$ is the triangular frequency response of the *i*th Mel-filter. These 24 Mel-filter bank energies are then transformed into 13 MFCC using DCT.

E. DCT

This is the process to convert the log Mel spectrum into time domain using DCT. The result of the conversion is called Mel Frequency Cepstral Coefficient. The set of coefficient is called acoustic vectors.

$$c_j = \sum_{i=1}^{24} \log(M_{FB}(i)) \sqrt{\frac{2}{24}} \cos(\pi j \frac{i-0.5}{24}), j \in (0, 12) \quad (7)$$

F. Data energy and Spectrum

The acoustic signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is need to add features related to the change in cepstral features over time. 13 feature (12 cepstral features plus energy).

$$Energy = \sum X^2[t] \quad (8)$$

Where X[t] = signal

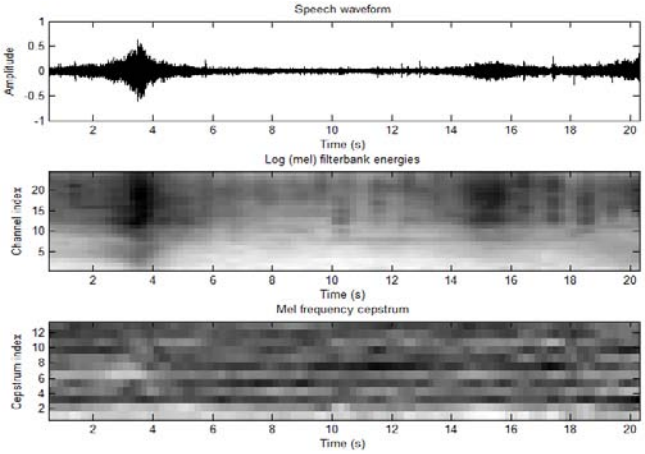


Fig. 8. Input Acoustic signal, corresponding log filterbank energies and Mel frequency cepstrum for low traffic density state

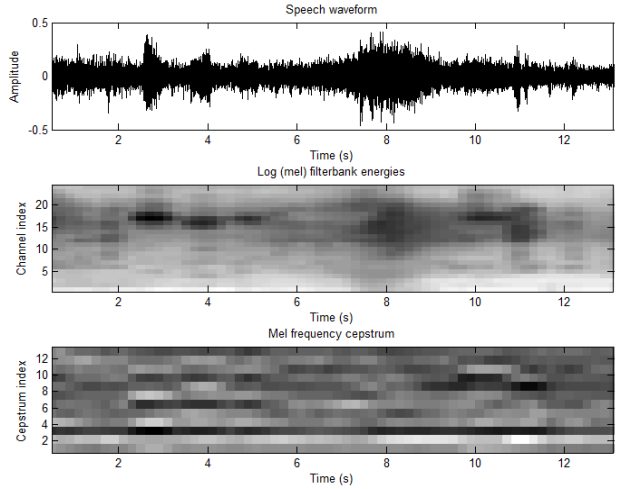


Fig. 9. Input Acoustic signal, corresponding log filterbank energies and Mel frequency cepstrum for Medium traffic density state

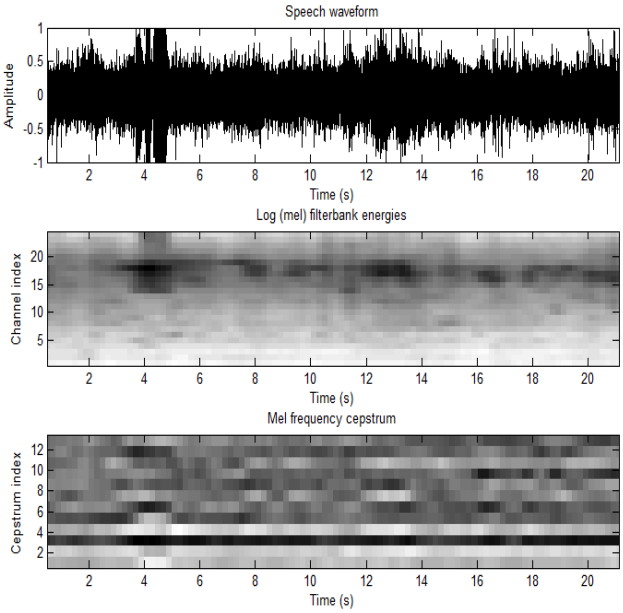


Fig. 10. Input Acoustic signal, corresponding log filterbank energies and Mel frequency cepstrum for Heavy traffic density state

5. Adaptive Neuro Fuzzy Classifier

Fuzzy classification assumes the boundary between two neighboring classes as a continuous, overlapping area within which an object has partial membership in each class [10]. Most of the classification problems consist of medium and large-scale datasets, example: genetic research, character or face recognition. For this different methods, such as neural networks (NNs), support vector machines, and Bayes classifier, have been implemented to solve these problems. The network-based methods can be trained with gradient based methods, and the calculations of new points of the network parameters generally depend on the size of the datasets. One of the network-based classifiers is the Neuro-Fuzzy Classifier (NFC) which combines the powerful description of fuzzy classification techniques with the learning capabilities of NNs.

A neural-fuzzy system is a combination of neural networks and fuzzy systems. The combination is such that the neural networks or neural networks algorithms are used to determine parameters of fuzzy system. This means, the main intention of neural-fuzzy approach is to create or improve a fuzzy system automatically by means of neural network methods. An adaptive network is a multi-layer feed-forward network where each node performs a particular function based on incoming signals and a set of parameters pertaining to node. Fuzzy classification systems, which are founded on the basis on fuzzy rules, have been successfully applied to various classification tasks [37]. The fuzzy systems can be constituted with neural networks, and resultant systems are called as Neuro-fuzzy systems [37]. The Neuro-fuzzy classifiers define the class distributions and show the input-output relations, whereas the fuzzy systems describe the systems using natural language. Neural networks are employed for training the system parameters in neuro-fuzzy applications. An ANFIS consist of input, membership function, fuzzification, defuzzification, normalization and output layers [37, 38, 39].

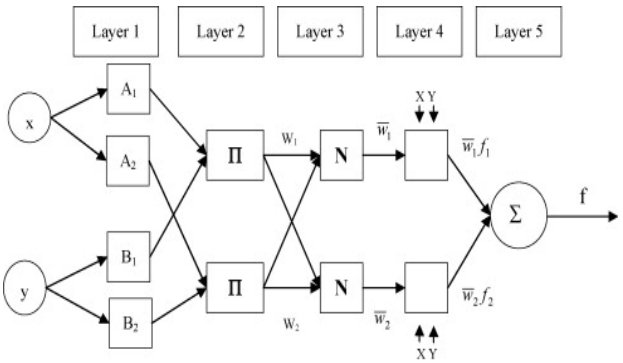


Fig. 11. An Adaptive Neuro-Fuzzy Classifier

Layer 1: Refer to Fig. 11, Every mode in this layer is an adaptive node with a node function where x (or y) is the input to node I and Ai (or Bi-2) is a linguistic label and $O_{1,1}$ is the membership grade of fuzzy set A(= A1, A2, B1or B2) and it specifies the degree to which the given input x (or y) satisfies the quantifier A. Usually, Gaussian membership functions are chosen to represent the linguistic terms.

$$\mu_A(x) = \text{Gaussian}(x; c, \sigma) = e^{\frac{-1}{2}(\frac{x-c}{\sigma})2} \quad (9)$$

$$O_{1,i} = \mu_{A_i}(x), \ for\ i = 1,2, or \quad (10)$$

$$O_{1,i} = \mu_{B_{i-2}}(y), \ for\ i = 3,4, \quad (11)$$

A Gaussian MF is determined complete by c and σ; c represents the MFs centre and σ determines the MFs width. In fact, any continuous, such as trapezoidal and triangular-shaped membership functions are also candidates for node functions in this layer.

Layer 2: Every node in this layer is a fixed node labeled Π, whose output is the product of all the incoming signals. Each node output represents the firing strength of a rule.

$$O_{2,i} = W_i = \mu_{A_i}(x)\mu_{B_i}(y),\ i = 1,2 \quad (12)$$

Layer 3: Every node in this layer is a fixed node labeled N. The ith node calculates the ratio of the ith rule's firing strength to the sum of all rules' firing strengths. Outputs of this layer are called normalized firing strengths.

$$O_{3,i} = \overline{W}_i = \frac{W_i}{W_1+W_2},\ i = 1,2 \quad (13)$$

Layer 4: Every node I in this layer is an adaptive node with a node function. Where $\overline{W}_i$ is a normalized firing strength from layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set of this node. Parameters in this layer are referred as consequent parameters.

$$O_{4,i} = \overline{W}_i f_i = \overline{W}_i(p_i x + q_i y + r_i), \quad (14)$$

Layer 5: The single node in this layer is a fixed node labelled ∑, which computes the overall output as the summation of all incoming signals. Overall output is:

$$O_{5,1} = \sum_i \overline{W}_i f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} \quad (15)$$

5.1 Scaled Conjugate Gradient (SCG)

Several gradient descent algorithms for feed-forward neural networks have poor convergence rate and depend on parameters which have to be specified by the user. In order to handle large-scale problems in an effective way, various optimization methods exists usually referred as Conjugate Gradient Methods. The Scaled Conjugate Gradient (SCG) algorithm is one of the Conjugate Gradient Method and is based on the second-order gradient supervised learning procedure [10]. In any iteration, the SCG computes two first-order gradients for the parameters to determine the second-order information. Two gradients are calculated per iteration of the SCG: the first gradient is calculated with a small step size, and the second gradient is calculated with a bigger step size. Experimental results by using SCG algorithm from [10] is as follows,

5.2 Experimental Results with ANFC_SCG

We have collected the road side cumulative acoustic signal samples from chhatrapati square to T-point of Nagpur city. Data were collected with 16 KHz sampling frequency. These data covered three broad traffic density classes (low, medium and heavy). Feature extraction is done using MFCC where primary window size is 500 ms and shift size is of 100 ms. When single feature frame is considered for the classification purpose we are getting average classification accuracy of 93.33% and when entire feature frames were considered, average classification accuracy increased to Approx 96% [Table 3].

Table 3. Classification accuracies of various traffic density classes based on single frame and Entire frames.

Traffic Density State	Single Frame	Entire Frames
Low	93.33 %	93.33 %
Medium	93.33%	96.67%
Heavy	93.33%	96.67%

Case 1: First 13 cepstral coefficients were considered.
Case 2: The entire feature vectors consisted of the first 13 MFCC coefficients are first obtained and then their first derivatives are computed.
Case 3: Second order derivatives of obtained first order derivatives in case 2 are computed.

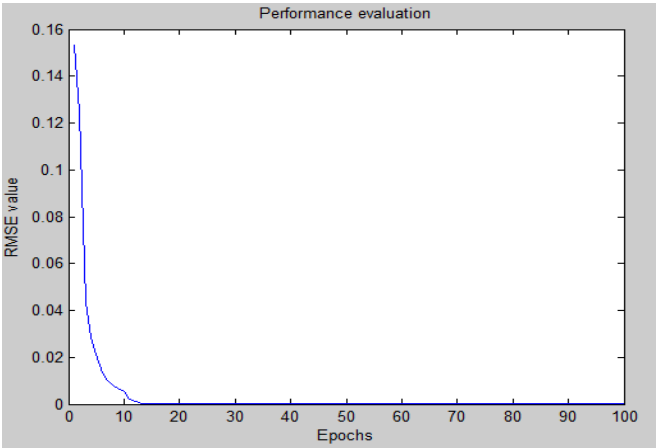


Fig. 12. Performance evaluation of recognition rate for case 1.

Table 4. Classification accuracies of various traffic density classes based on first and second order derivatives of first frame.

Traffic Density Class	First order derivative	Second order derivative
Low, Medium and Heavy	77.78 %	~75 %

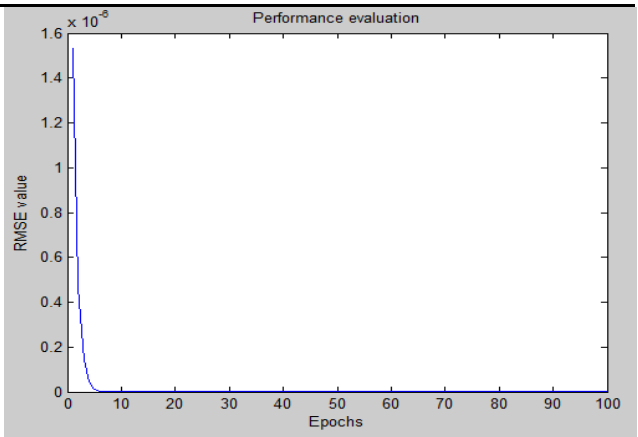


Fig. 13. Performance evaluation of recognition rate for case 2.

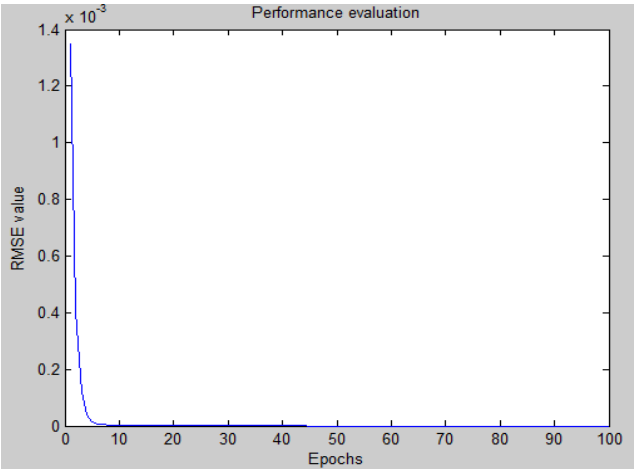


Fig. 14. Performance evaluation of recognition rate for case 3.

5.3 Adaptive Neuro Fuzzy Classifier with Speedup Scaled Conjugate Gradient (SSCG)

The training time of networks is a main problem for large-scale parameters, and computation cost of the SCG algorithm per iteration is more expensive for large-scale problem. The training NFC with SCG for large-scale problems could consume more time, such as days or weeks, with any personal computer.

The speeding up scaled conjugate gradient algorithm proposed by Bayram Cetisli and Atalay Barkana [39] shortens the training time per iteration of SCG without affecting the convergence rate of the training. Step wise execution and details of the SSCG algorithm in presented in [39]. Training time can be sufficiently reduced by using SSCG for large scale problems. Moreover our focus in on classification purpose only because gathered data set is small, however if feature dataset is very high for various traffic density states then in that case SSCG will certainly be best option.

5.4 Experimental Results with ANFC-SSCG

Table 4: Average Classification Accuracy of Various Traffic Density States using ANFC_SSCG (in %) for the features extracted using MFCC

Traffic Density State	13 Coefficients	Entire Coefficients
Low	93.33 %	93.33 %
Medium	93.33%	96.67%
Heavy	93.33%	96.67%

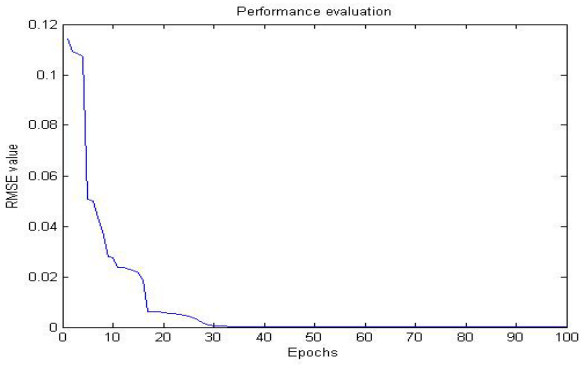


Fig. 15. Performance evaluation of ANFC_SSCG when applied on features extracted using MFCC

The classification accuracy of all the three classes' increases as the primary analysis window size is increased from 40 to 500 ms and shift size from 20 to 100 ms. (Accuracy obtained with lower primary window and shift is not included as part of this paper). This is due to the fact that the traffic density class (state) is a slow-changing physical process, owing to its inherent physical constraints. (Speeds of the vehicles are bounded between (0, 100) km/h, and hence, the density of the traffic cannot change at an arbitrarily high rate.) On average, the traffic density class (state) on a particular road segment can be expected to change from, for example, jammed to medium-flow and medium-flow to free-flow on a time scale of 5–30 min or even higher. This is unlike speech signals, where a phoneme (a basic classification/recognition unit) can change over a 20–80 ms time period. Furthermore, the spectral characteristics of the cumulative acoustic signal do not significantly change over the time spans of about 200 ms. Therefore, a primary analysis window of size 200 or 500 ms and a window shift of 100 ms seem to be a reasonable choice

6. Conclusion

This paper describes a simple technique which uses MFCC features of road side cumulative

acoustic signal to classify traffic density state as Low, Medium and Heavy using Adaptive Neuro-Fuzzy Classifier. As this technique uses simple microphone (cost: 500 Rs) so its installation, operational and maintenance cost is very low. This technique work well under non lane driven and chaotic traffic condition, and is independent of lighting condition. Classification accuracy achieved using Adaptive Neuro-Fuzzy classifier is of 93.33% for 13-D MFCC coefficients and approx 96% when entire features were considered, 77.78% for first order derivatives and ~75% for second order derivatives of cepstral coefficients. From the results it is observed that first and second order derivatives are not as much relevant but may be 13 D coefficients and their 1st and 2nd derivatives, together combine 39-D coefficients will improve the accuracy.

The research on vehicular acoustic signal which is mixture of engine noise, tyre noise, noise due to mechanical effects etc. expands from vehicular speed estimation to density estimation. The use of road side acoustic signal seems to be an alternative, research shows acceptable accuracy for acoustic signal. Vehicular classification with Acoustic signals proved to be excellent approach particularly for battlefield vehicles, and also for city vehicles.

Clearer definitions of scenarios and applications are required to generate a more consistent body of work. New application areas are likely to emerge for traffic signal timings optimization using cumulative acoustic signals and also classification of motorcycles proved to be emerging area for research. Finally the classification systems can be extended in a way that extracted features are utilized as characteristic fingerprints, which affords tracking of vehicles over multiple sensor nodes.

References

[1] Chen Xiao-feng, Shi Zhong-ke and Zhao Kai, "Research on an Intelligent Traffic Signal Controller," 2003 IEEE.

[2] Chunxiao LI and Shigeru SHIMAMOTO, "A Real Time Traffic Light Control Scheme for Reducing Vehicles CO2 Emissions," The 8th Annual IEEE Consumer Communications and Networking Conference - Emerging and Innovative Consumer Technologies and Applications.

[3] R. Lopez-Valcarce, C. Mosquera, and R. Perez-Gonzalez, "Estimation of road vehicle speed using two omnidirectional microphones: A maximum likelihood approach," EURASIP J. Appl. Signal Process., pp. 1059–1077, 2004.

[4] Autoscope. [Online]. Available: http://www.autoscope.com

[5] Citilog. [Online]. Available: http://www.citilog.com

[6] CRS, Computer Recognition Systems. [Online]. Available: http://www.crs-traffic.co.uk

[7] Ipsotek. [Online]. Available: http://www.ipsotek.com/

[8] Traficon. [Online]. Available: http://www.traficon.com

[9] Sun CT, Jang JSR, "A neuro-fuzzy classifier and its applications," In Proceedings of IEEE International Conference on Fuzzy Systems, San Francisco 1:94–98, 1993.

[10] Martin F. Møller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning," Neural Network 6(4):525–533. 1993.

[11] S. A. Amman and M. Das, "An efficient technique for modeling and syn-thesis of automotive engine sounds," IEEE Trans. Ind. Electron., vol. 48, no. 1, pp. 225–234, Feb. 2001.

[12] U. Sandberg, "Tyre/road noise—Myths and realities," in Proc. Int. Congr. Exhib. Noise Control Eng., The Hague, Netherlands, Aug. 27–30, 2001.

[13] Road Directorate-Ministry of Transport, "Noise Reducing Pavement," Road Directorate, Danish Road Institute Tech Report 141, Apr 2005.

[14] U. Sandberg and A. J. Ejsmont, "Tyre/Road Noise Reference Book," Kisa, Sweden: Infomex, 2002, SE-59040.

[15] R. A. G. Graf, C. Y. Kuo, A. P. Dowling, and W. R. Graham, "On the horn effect of a tyre/road interface—Part I: Experiment and computation," Journal of Sound and Vibration, vol. 256, pp. 417–431, 2002.

[16] C. Y. Kuo, R. A. G. Graf, A. P. Dowling, and W. R. Graham, "On the horn effect of a tyre/road interface—Part II: Asymptotic theories," Journal of Sound and Vibration, vol. 256, pp. 433–445, 2002.

[17] V. Cevher, R. Chellappa and J. H. McCllelan, "Vehicle Speed Estimation Using Acoustic Wave Patterns", IEEE Trans. on Signal Processing, Vol. 57, No. 1, Jan 2009.

[18] J. G. Lilly, "Engine Exhaust Noise Control, " [Online] Available: http:www.ashraeregion7.org

[19] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," Proc. IEEE, vol. 87, pp. 973–973, 1999.

[20] R. E. Eskridge, and J. C. R. Hunt, "Highway Modeling. Part I: Prediction of Velocity and

Turbulence Fields in the Wake of Vehicles," Amer. Meteorolog. Soc., Vol. 79, pp. 387-400, 1979.

[21] N. Sarigul-Klijn, D. Dietz, D. Karnopp, and J. Dummer, "A computational Aeroacoustic Model for Near and Far Field Vehicle Noise predictions," New York: The Amer. Inst. Aeronaut. Astronaut., 2001.

[22] D. I. Robertson and R. D. Bretherton, "Optimizing networks of traffic signals in real time—The SCOOT method," IEEE Transaction on Vehicular Technology, vol. 40, no. 1, pp. 11–15, Feb. 1991.

[23] B. G. Quinn, "Doppler speed and range estimation using frequency and amplitude estimates," *J. Acoust. Soc. Amer.*, vol. 98, no. 5, pp. 2560–2566, Nov. 1996.

[24] C. Couvreur and Y. Bresler, "Doppler-based motion estimation for wide-band sources from single passive sensor measurements," in *Proc. IEEE ICASSP*, Apr. 1997, pp. 21–24.

[25] S. Chen, Z. P. Sun, and B. Bridge, "Automatic traffic monitoring by intelligent sound detection," Proc. IEEE Intelligent Transportation Systems Conf., Nov. 1997.

[26] S. Chen and Z. P. Sun, "Traffic sensing by passive sound detection," in Proc. Sensors and Their Applications VIII Conf., Glasgow, Scotland, U.K., Sept. 1997

[27] S. Chen, Z. Sun, and Bryan Bridge, "Traffic Monitoring Using Digital Sound Field Mapping," IEEE Transactions on vehicular technology, vol. 50, no. 6, pp.1582-1589, November 2001

[28] K. W. Lo and B. G. Ferguson, "Broadband passive acoustic technique for target motion parameter estimation," IEEE Trans. Aerosp. Elect. Syst., vol. 36, pp. 163–175, 2000.

[29] V. Cevher, R. Chellappa and J. H. McCllelan, "JOINT ACOUSTIC-VIDEO FINGERPRINTING OF VEHICLES, PART I", in Proc. Of ICASSP, II- 745-748, IEEE 2007.

[30] J. Kato, "An Attempt to Acquire Traffic Density by Using Road Traffic Sound," Active Media Tech., pp. 353-358, IEEE 2005.

[31] S. Sampan, "Neural fuzzy techniques in vehicle acoustic signal classification," Ph.D. dissertation, Dept. Elect. Eng., Virginia Poly. Inst. State Univ., Blacksburg, VA, 1997.

[32] G. Succi, T.K. Pedersen, R. Gampert, G. Prado, "Acoustic Target Tracking and Target Identification - Recent Results," Proc. SPIE, Vol. 3713, pp. 10-21, 1999.

[33] A.Y. Nooralahiyan, H.R. Kirby, "Vehicle Classification by Acoustic Signature," Mathl. Comput. Modelling, Vol. 27, No. 9-11, pp. 205-214, 1998.

[34] X. Wang, H. Qi, "Acoustic target classification using distributed sensor arrays," Proc. IEEE ICASSP, Vol. 4, pp. 4186-4189, 2002.

[35] S. Theodoridis, K. Koutroumbas, "Pattern Recognition", San Diego, Elsevier, Academic Press, 2006

[36] H. Choe, R. Karlsen, G. Gerhart, and T. Meitzler, "Wavelet based ground vehicle recognition using acoustic signals," in Proc. SPIE, 1996, vol. 2762, pp. 434–445.

[37] Jang, J.S.R., Sun, C.T. and Mizutani E., "Neuro-fuzzy and soft computing", Upper Saddle River :Prentice Hall, 1997.

[38] Cetisli, B., "Development of an adaptive neuro-fuzzy classifier using linguistic hedges: Part 1," Expert Systems with Applications, doi:10.1016/j.eswa.2010.02.108, 2010.

[39] Bayram Cetis and Atalay Barkana, "Speeding up the scaled conjugate gradient algorithm and its application in neuro-fuzzy classifier training," Soft Comput (2010) 14:365–378, DOI 10.1007/s00500-009-0410.

[40] A. Jazayeri, H. Cai, J. Y. Zheng, and M. Tuceryan, "Vehicle detection and tracking in car video based on motion model," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 2, pp. 583–595, Jun. 2011.

[41] A. Faro, D. Giordano, and C. Spampinato, "Evaluation of the traffic parameters in a metropolitan area by fusing visual perceptions and CNN processing of webcam images," IEEE Trans. Neural Netw., vol. 19, no. 6, pp. 1108–1129, Jun. 2008.

[42] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya, "Real-time measurement of link vehicle count and travel time in a road network," IEEE Trans. Intell. Transp. Syst., vol. 11, no. 4, pp. 814–825, Dec. 2010.

[43] R. Cucchiara, M. Piccardi, and P. Mello, "Image analysis and rule based reasoning for a traffic monitoring system," IEEE Trans. Intell. Transp. Syst., vol. 1, no. 2, pp. 119–130, Jun. 2000.

[44] S. Kamijo, Y. Matsushita, and K. Ikeuchi, "Traffic monitoring and acci-dent detection at intersections,"IEEE Trans. Intell. Transp. Syst.,vol.1, no. 2, pp. 108–118, Jun. 2000.

[45] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," Transp. Res. C, Emerging Technol., vol. 6, no. 4, pp. 271–288, Aug. 1998.

[46] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphone," inProc. SenSys, Nov. 2008, pp. 323–336.

[47] J. Lee and A. Rakotonirainy, "Acoustic hazard detection for pedestrians with obscured hearing," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1640–1649, Dec. 2011.

[48] Vivek Tyagi, Shivkumar Kalyanaraman, and Raghuram Krishnapuram, "Vehicular Traffic Density State Estimation Based on Cumulative Road Acoustics," IEEE Transactions on Intelligent Transportation Systems. 2012.

[49] R. Cucchiara, M. Piccardi, and P. Mello, "Image analysis and rule based reasoning for a traffic monitoring system," IEEE Transaction on Intelligent. Transp. Syst., vol. 1, no. 2, pp. 119–130, Jun. 2000.

[50] S. Kamijo, Y. Matsushita, and K. Ikeuchi, "Traffic monitoring and accident detection at intersections," IEEE Transaction on Intelligent Transp. Syst.,vol.1, no. 2, pp. 108–118, Jun. 2000.

[51] R. Sen, B. Raman, and P. Sharma, "Horn-ok-please," in Proc. ACM MobiSys, San Francisco, CA, 2010, pp. 137–150.

[52] R. O. Duda, P. E. Hart, D.G. Stork, "Pattern Classification," John Wiley & Sons LTD, 200