TRAFFIC DENSITY STATE ESTIMATION BASED ON ACOUSTIC FUSION

Vikas Joshi, Nithya Rajamani, Naveen Prathapaneni, L. V. Subramaniam

IBM India Research Labs, New Delhi
email: {vijoshi7, nitrajam, n.prathapaneni, lvsubram}@in.ibm.com

ABSTRACT

In this paper, we propose an acoustic fusion based approach to classify the traffic density states. In particular, we combine the information from mel-frequency cepstral coefficients (MFCC) based classifier, which models the cumulative road side signal and honk event based classifier. Honk based classifier is obtained by modeling the honk statistics for each traffic class, *viz*., Jam, Medium and Free. We study in detail the discriminative capabilities of honk information based classifier. Decisions from MFCC and honk classifier are then combined in probabilistic framework with an appropriate fusion strategy. We also propose to use prior honk information in-order to further improve the classification results. Classification results show good performance even with $10s$ of audio data.

***Index Terms*—** Traffic state detection, Acoustic modeling, fusion, honks, MFCC

1. INTRODUCTION

Traffic congestion is an important problem around the world and is more rampant in the developing regions like South-East Asia. Knowing the traffic conditions at all the locations could help authorities to regulate the traffic flow. Magnetic loop detectors[1] are used in several developed nations to sense the traffic. These techniques involve a heavy implementation cost and are also not readily applicable in the developing regions, because of non-lane based and chaotic traffic conditions. A video showing chaotic traffic conditions in developing regions is present in [2]. Video imaging based techniques have been proposed for traffic sensing[3], however they too have similar limitations of cost and assumptions of orderly traffic conditions. Recently acoustics based traffic sensing techniques have been proposed [4] [5]. Acoustic sensor based approach is lucrative because of its ability to sense traffic states even in chaotic traffic conditions and also due to its low implementation cost. Road side cumulative acoustic signal is a mixture of vehicular noise mainly consisting of tire noise, engine noise, air turbulence, and honks[6]. Acoustic signals are distinctly different for different traffic conditions namely *Jam* ($0 - 10kmpH$)*, Medium* ($10 - 40kmpH$) *and Freeflow* ($> 40kmpH$). Free condition is dominated by air-turbulence and tire noise, while jam condition has engine idling noise and honks.

Vivek *et al*.[4] proposed MFCC based traffic state classifier. Cumulative signal was characterized using MFCC features and modeled using Gaussian mixture models (GMM). In [4] data was recorded in a controlled setting with an omni-directional microphone, while we collected in a more general setting using smart phones, so that data can be obtained by a single fixed sensor and even by participative sensing[7]. Although impressive results were shown for MFCC classifier in [4], we obtained relatively poor results with data collected in more general settings using smart phones. Hence there was a scope and need for further improvement in the classification results. During the data collection process, it was observed that along with speed of vehicles, honks were also indicative of traffic condition. This motivated us to use the honk information along with information captured by MFCC classifier to make more accurate decisions. Although MFCC classifier models the cumulative signal (which contain honks), explicit modeling of honk information is not done. Honk classifier is built by modeling the number of honks observed for each class. In this paper, we study in detail, the characteristics of honk information based classifier and propose a fusion classifier, combining the decision of MFCC and honk classifiers alone. We use probabilistic framework to combine the decision of MFCC and honk classifier. Classification accuracy of fusion classifier is better than MFCC and honk classifiers. The classification accuracy was further improved by using prior honk information, even with just $10s$ of acoustic data.

Since acoustic signals for three traffic states had distinctly different spectral content[4], MFCCs were used as basic parameterization, which capture the spectral shape of the signal. In this paper, MFCC classifier is implemented as explained in [4]. The short time Fourier transform (STFT) of acoustic signal is obtained by windowing the signal with the window size of $100ms$. It is then passed through a filter-bank followed by log compression. Finally 13 dimensional coefficients are obtained by applying discrete cosine transform on the log filter-bank coefficients. Delta and delta-delta coefficients are appended to obtain final feature vector of 39 dimensions. The next frame is obtained by shifting the window with a shift size of $50ms$. GMMs were used to model the MFCC features and were shown to perform well in [4]. GMMs also easily provide soft decision and hence is more suitable (than discriminative classifiers) for fusion approach.

1.1. Relation to prior work

There have been few recent approaches proposed for using audio data in traffic sensing [4] [8] [9] [10]. In our proposed approach, we explicitly model the honk information along with MFCC based classifier proposed in [4]. Rijurekha *et al.* [8][5] proposed a two-sensor based architecture using honks signals to estimate traffic state. Our approach is different from [5] in terms of architecture (requires only single sensor) and also in approach of modeling the sources of information. Approach in [5] heavily relied on the honks, while in our proposed approach the final decision is based on weighted likelihoods of MFCC and honk classifier, weighted according to the confidence in the each classifier. In [10] acoustic sensor network is used to estimate the speed of vehicle by calculating acoustic time delay from one sensor to other. However, in a chaotic traffic conditions,

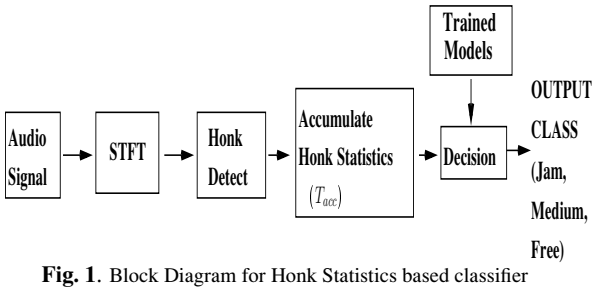


Fig. 1. Block Diagram for Honk Statistics based classifier

this approach may not perform well due to interference from other noises, which is not well addressed. In [9], a combination of smart phone sensors are used which also use audio as one of sensor inputs.

The rest of the paper is organized as follows: In section 2, system architecture is explained, followed by honk based classifier in section 3. Section 4 contains fusion strategy followed by experimental results in section 5 and conclusion in section 6.

2. SYSTEM ARCHITECTURE AND DATA COLLECTION

The system architecture consists of a single microphone installed road-side to record the audio data and a transmitter to send the audio data along with the location information to a centralized server. Data can also be sent by participatory sensing with people uploading the data using smart phones to the central server[7]. All the processing and classification is done at the central server. Poller checks for the data and traffic state is updated for every one minute. Even though decision is made at every minute, whole one minute audio might not be available due to latency and bandwidth cost constraints. $320k$bytes of data need to be transmitted to send $10s$ of data sampled at $16kHz$. Using larger audio would also discourage participative sensing. Hence we attempt to get improved results with as small as $10s$ of audio data. The sample demo describing the architecture is present in [2]. The current architecture receives the data that is uploaded from the smart phones. Fixed sensors on the road-side are yet to be deployed.

Data Collection: Since the fixed sensors are not yet deployed, audio data was recorded from Samsung Galaxy S2 smart phone. Data was collected from two cities in India, namely New Delhi and Hyderabad. Approximately 3 hours of data was collected from Delhi with 1 hour of each traffic state (jam, medium and free) and the sampling rate was $16k$Hz. Around 1.5 hours of traffic data was collected from Hyderabad location, with $\sim$ 30 minutes of each traffic state. The entire data-set and details about the data collection is present in [2]. Delhi data was divided into train and test sets and are referred as DL-Train and DL-Test respectively. Hyderabad data is used only for testing and is referred as HD-Test.

3. HONK STATISTICS BASED CLASSIFIER

Number of honks at a particular location could provide useful information about the traffic state. In general, more number of honks (observed over a certain time interval) would correspond to jam condition. Although honking would depend on the attributes of the driver and location of driving, a more chaotic condition would naturally provoke a tendency and the need to honk. Hence we wish to leverage honk information and study the discriminative capability present in the honk information. In this section we describe a honk statistics based classifier for traffic state estimation. The term *Honk Statistics* mentioned in the paper correspond to percentage of honk frames in a defined time interval. Honks have been previously used in [5] [9] as one of the feature vector in their discriminative classifiers. While in this paper, we build separate classifier by modeling the honk information using generative models.

Fig. 1 shows the block diagram of honk statistics based classifier which essentially involves a training and testing phase. The training steps are as explained below:

1. **Short time Fourier Transform (STFT)**: Audio signal is divided into frames with window size of $100ms$ and shift size of $50ms$ as used for MFCC classifier. FFT is then applied on the windowed signal.
2. **Honk detection**: Honks frames are detected from the STFT of the audio signal. In section 3.1 two honk detection algorithms are discussed.
3. **Accumulation of honk statistics over defined time interval**: Percentage of the honks frames (honk statistics) within a defined time interval is calculated. This time interval is referred to as *accumulate time interval* denoted by T_{acc}. For $T_{acc} = 10s$, there are 200 frames (with frame shift size of $50ms$). Honk detection is done for each frame. Percentage of honk frames is then calculated from 200 frames which corresponds to $T_{acc} = 10s$. Approach to decide the T_{acc} is explained later in the section 3.3.
4. **Model the traffic class**: Models are built for each traffic class with percentage of honks as the discriminating feature. Percentage of honks is calculated over time T_{acc}. Model parameters change significantly with the different choice of T_{acc}. Hence it is important to choose an appropriate T_{acc}.

Testing phase is explained in the Fig. 1. Given the audio signal, honk statistics are obtained for the defined time interval (T_{acc}) as explained in training steps and in Fig. 1. Each class likelihoods are then calculated using the respective trained models. Finally, decision is made based on the maximum likelihood approach for every T_{acc} secs.

In the following subsections we describe honk detection algorithm followed by choice of class densities to model the traffic states. Finally in section 3.3 classification results are reported along with the approach to choose T_{acc}.

3.1. Honk Detection Algorithms

Honk frames are typically characterized by number of harmonic peaks in the frequency range of $2kHz$ to $4kHz$, referred to as honk frequency range. Peak to Average amplitude approach was discussed in [8], however it was sensitive to spurious honks. Here we also investigate Variance based approach which we find is more robust to spurious honks.

- PeakvsAvg: This approach was explained in [8] (referred as PeakvsAvgAllFreq) where ratio of Peak amplitude to Average amplitude is compared with a threshold (T) to detect the honk frame. Peak amplitude is calculated within frequency range of $2kHz$ to $4kHz$ and average amplitude is calculated as sample mean over all the frequency components.
- Variance based approach: Honks are characterized by multiple peaks within the frequency range of $2kHz$ to $4kHz$ (honk frequency range). Since there are multiple peaks, variance of the squared magnitude values of the frequency spectrum is high. Hence the *variance of the amplitudes of squared magnitude frequency spectrum*, within the honk frequency range is compared with a threshold to detect the honk frames.

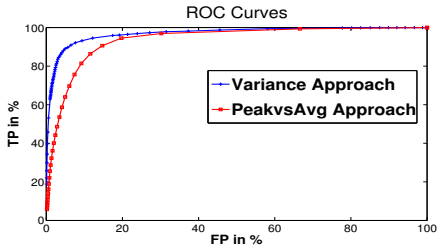


Fig. 2. ROC curves for Variance and PeakvsAvg approach

3.1.1. Comparison of the Honk Detection Algorithms

The above two honk detection algorithms are tested on the DL-Train data with approximately $90mins$ of traffic data. $90mins$ dataset contains approximately $30mins$ data from each traffic state. Ground truth is manually labeled using the Audacity tool[11]. The dataset along with the ground truth labels can be found at [2]. Fig. 2 shows the ROC curve for both approaches. It can be seen that Variance approach outperforms PeakvsAvg approach. Along with detecting the honks, the above two algorithms are also tested for their ability to discriminate between the traffic states. In the following sections we evaluate both the algorithms for classification results.

3.2. Choosing the pdf to model traffic states

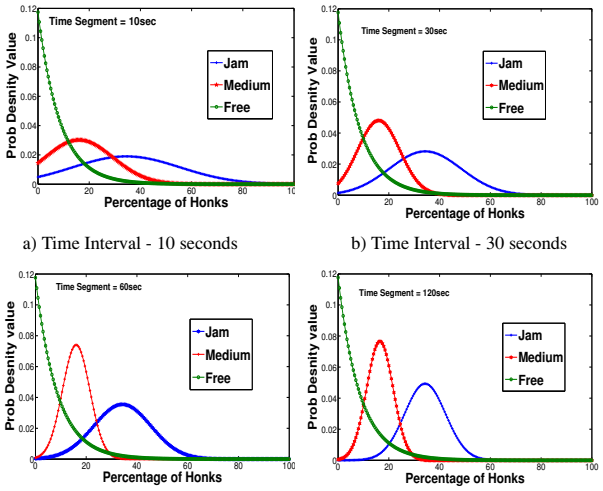
Suitable pdfs are chosen depending upon the distribution of the honk statistics for the each class. In case of free traffic class, it was observed that honk statistics were more concentrated near the origin and exponentially decreased as moved further. Hence exponential density function was chosen to model the free class. Histogram of the honk statistics for medium class showed an increase in the histogram count as the percentage of honks increased and then it decreased gradually. Hence Gaussian density was used to model medium class. According to heuristics, the likelihood of the jam class should increase as the number of honks is increased. However, Gaussian density is used to model jam class, as training data showed Gaussian distribution.

3.3. Choice of accumulation time interval (T_{acc}) and classification results

Accuracy of the honk statistics classifier largely depends on the time interval over which the honk statistics are calculated. A very small time interval ($<$ 10secs) would not provide enough evidence to differentiate between the three classes. A large time interval could also give errors since the traffic state could change within the specified time interval. Hence an appropriate time interval need to be chosen which has enough discriminative evidence and also the traffic state does not change within the specified time interval. However, architecture of the overall system and cost of the audio data could constrain the choice of time interval. Figs. 3(a), 3(b), 3(c) and 3(d) show the pdf of the honk statistics for time intervals of $10s, 30s, 60s$ and $120s$ respectively. It can be seen that the discrimination between the class densities increase as the time interval is increased from $10s$ to $120s$.

T_{acc}	Variance Approach				PeakvsAvg Approach			
	Jam	Med	Free	Overall	Jam	Med	Free	Overall
10	69.4	40.8	94.7	67.6	50.4	40.8	92.9	65.5
30	81.93	56.92	100	79.58	53.1	49.5	94.7	65.1
60	85.3	67.4	100	83.4	60	49.5	98.2	75.1
120	89.1	82.1	100	89.5	64.3	42.8	100	68.3

Table 1. Classification accuracy of honk based classifier using Variance based approach and PeakvsAvg approach



a) Time Interval - 10 seconds b) Time Interval - 30 seconds

c) Time Interval - 60 seconds d) Time Interval - 120 seconds

Fig. 3. Honk Statistics (Percentage of honks) in different time intervals for Variance based approach. It can be seen that discrimination between the three classes increases as the time interval is increased from 10sec to 120sec (Assumption: Traffic state does not change within the defined time interval).

Table 1 shows the classification results for both approaches for different accumulation time (T_{acc}). DL-Train is used for training and testing is done with DL-Test and HD-Test data. It can be seen that Variance approach out-performs PeakvsAvg approach. As expected (from Fig. 3), accuracies improve as T_{acc} is increased.

4. FUSION CLASSIFIER

The goal of the fusion classifier is to constructively combine the information from the MFCC and honk based classifier. It is important that poorly performing classifier do not skew the final decision. Quite a few approaches have been proposed in the literature to implement the fusion strategy [12][13] [14]. In our case, fusion is done at the decision level[15] by a simple weighted based fusion approach. The probability of each class obtained from MFCC and honk classifiers are weighted with the confidence measure of the each classifier. The probability of the feature vector X, belonging to an event $e_i \in Jam, Medium, Free$ is given by,

$$P(X = e_i) = P(X|C^M_{e_i})\gamma_{e_i} + (1 - \gamma_{e_i})P(X|C^H_{e_i}) \quad (1)$$

where $P(X|C^M_{e_i})$ and $P(X|C^H_{e_i})$ represent the probability of X belonging to event e_i obtained from MFCC and honk classifiers respectively. γ_{e_i} and $1 - \gamma_{e_i}$ are the relative confidence measures of the MFCC and honk based classifiers for event e_i. The probabilities are obtained using the model likelihoods by,

$$P(X|C^M_{e_i}) = \frac{L(X = e_i|\lambda^M_{e_i})}{\sum_i L(X = e_i|\lambda^M_{e_i})} \quad (2)$$

where $L(X = e_i|\lambda^M_{e_i})$ is the likelihood of the event e_i w.r.t to MFCC model λ_{e_i} for event e_i. Same formula is used for the honk

T_{seg}	MFCC				Honk ($T_{acc} = T_{seg}$)				Fusion			
(s)	Jam	Med	Free	Overall	Jam	Med	Free	Overall	Jam	Med	Free	Overall
10	75.19	86.25	97.64	**86.36**	69.43	40.80	94.73	**67.63**	87.62	83.22	97.64	**89.49**
20	81.89	93.70	98.82	**91.47**	78.19	56.40	97.64	**77.41**	89.03	91.89	98.82	**93.25**
30	86.19	96.20	100	**94.13**	81.83	56.92	100	**79.58**	94.62	96.20	100	**96.94**

Table 2. Recognition accuracies(%) for the traffic classes for MFCC, Honk and Fusion classifiers

T_{acc}	MFCC (Overall)	Fusion with Prior information($T_{seg} = 10s$)			
(s)	$T_{seg} = 10s$	Jam	Med	Free	Overall
10	**86.36**	87.62	83.22	97.64	**89.49**
30	**86.36**	87.22	87.82	98.23	**91.09**
60	**86.36**	91.66	88.46	98.23	**92.78**
120	**86.36**	93.88	88.46	98.82	**93.72**

Table 3. Recognition accuracy(%) for fusion approach using prior honk statistics

based classifier as well. There are few confidence measures proposed in [15] [13]. In our case, we found that weighting with the recognition accuracy of respective classifier obtained from the training data gave best results. Thus γ_{e_i} is obtained as

$$\gamma_{e_i} = \frac{R_M(e_i)}{R_M(e_i) + R_H(e_i)} \quad (3)$$

where $R_M(e_i)$ and $R_H(e_i)$ are the recognition accuracies for event e_i obtained on the training data (DL-Train) for MFCC and honk classifier respectively.

4.1. Modified Fusion strategy

Table 1 shows that honk classier has very low classification accuracy when small amount of audio data ($10s$) is available. However the MFCC classifier was seen to perform well even with $10s$ of audio data. Hence fusion classifier decision could be skewed by the low performing honk classifier. Figs. 3(a) and (b) show a lot of overlap between the class densities at low percentage of honks ($< 40\%$ of honks). However, if the percentage of honks is greater than 40% then overlap is significantly less. Thus honk classifier decision is reliable if percentage of the honks is $> 40\%$. A simple modification is done on fusion approach when less than $30s$ of audio data is available. The modification is as follows:

- If percentage of honks $< 40\%$, use only MFCC classifier.
- If percentage of honk $>= 40\%$, use Fusion classifier as defined by the Eqn. 1.

5. EXPERIMENTAL SETUP AND RESULTS

MFCC classifier is implemented as explained in [4]. 39 dimensional cepstral features were used as mentioned in section 1. The window size and shift size were chosen to be $100msec$ and $50msec$ respectively. A larger window size would provide better results [4], however require a large computation time (FFT calculation time increases) resulting in a larger latency. Three GMMs with 11 mixtures were built (one for each class) with 30 mins of data for each model. Honk classifier and fusion strategy is as explained in section 3 and section 4 respectively.

During the testing process, decision is made at every fixed time segment (T_{seg}). Results are presented for 3 time segments of $10s$, $20s$ and $30s$. Likelihoods are calculated for each frame ($100ms$) and average of the likelihoods is calculated over the defined T_{seg}, for each class. Maximum average likelihood criteria is chosen to make the decision. Here honk accumulation time $T_{acc} = T_{seg}$. The recognition results are shown in the Table 2. DL-Train data was used for training. Test set included DL-Test and HD-Test data-set as mentioned in section 2. Classification accuracies are shown for MFCC, honk based and fusion classifier. MFCC based classifier always performs better than honk classifier. It can be seen that the fusion approach outperforms the MFCC and honk based classifier consistently at all time segments.

5.0.1. Fusion with prior honk statistics

The honk classifier accuracy increases as accumulation time T_{acc} is increased as seen from the Table 1. Hence previous time segment honk statistics can be used by the honk classifier to make more accurate decisions. MFCC classifier uses only current $10s$ of acoustic data to make the decision, hence MFCC classifier is the representative of the current acoustic data that was sent by the transmitter. As a result, the classification accuracy of the MFCC classifier will be same as reported in the Table 2. Honk based classifier will use the honk information stored from previous time segment, along with the current segment data. This prior information improves the honk classifier accuracy and thereby the fusion results are also improved as shown in Table 3. Decision is done at every $10s$ i.e., $T_{seg} = 10s$, while the honk classifier uses prior information corresponding to different T_{acc} as shown in Table 3.

5.1. Future Work - Sensitivity of classifier to mismatched data conditions

Both MFCC and honk classifiers are seen to be sensitive to mismatch in data conditions resulting in degradation of the results. Some of the notable sources of mis-match are a) Width of the road b) Type of vehicles c) Pedestrian interference d) Type of recording device. Hence models trained on particular type of road segment cannot be used on road segment having different characteristics as mentioned above. This was also reported in [5] where location specific models were used. In future work we look to address this problem by adapting the models with the location specific data.

6. CONCLUSIONS

An approach to traffic density state classification is proposed, using road side collected acoustic data. The proposed approach and architecture is well suited in developing regions having chaotic traffic conditions. Motivation and algorithm is in accordance with real-time deployment with data being collected either from fixed road-side placed sensor or from participatory sensing. Along with modeling cumulative signal using MFCC features, we also exploit honk events to improve the classification accuracy. Honk based classifier is studied and Variance approach used for honk detection is seen to perform well. It is also seen that honk statistics became more discriminative as accumulation time (T_{acc}) is increased. Hence honk statistics were stored and used as prior information for processing next time segment, which significantly improved the classification results.

7. REFERENCES

[1] D Robertson and Bretherton David, "Optimizing networks of traffic signals in real time the scoot method," *IEEE Transactions on Vehicle Technology*, vol. 40, 1991.

[2] ," http://kalpa.haifa.il.ibm.com:9080/AcTraf/downloads.

[3] Li Li, Chen Long, Huang Xiaofei, and A Jian Huang, "Traffic congestion estimation approach from video using time-spatial imagery," in *International Conference on Intelligent Networks and Intelligent Systems*, 2008.

[4] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular Traffic Density State Estimation Based on Cumulative Road Acoustics," *IEEE Transactions on Intelligent Transportation Systems*, , no. Sept, 2012.

[5] Rijurekha Sen, Pankaj Siriah, and Bhaskaran Raman, "RoadSoundSense : Acoustic Sensing based Road Congestion Monitoring in Developing Regions," *SECON*, pp. 125–133, 2011.

[6] S.A. Amman and M. Das, "An efficient technique for modeling and synthesis of automotive engine sounds," *IEEE Transactions on Industrial Electronics*, vol. 48, no. 1, pp. 225–234, 2001.

[7] Raghu Ganti, Fan Ye, and Hui Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[8] Rijurekha Sen, Bhaskaran Raman, and Prashima Sharma, "Horn-Ok-Please," in *ACM Mobisys*, San Fransico, USA, 2010.

[9] Prashanth Mohan, Venkata N, and Ramachandran Ramjee, "Nericell : Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones," *ACM Sensys*, 2008.

[10] Barbara Barbagli, Gianfranco Manes, Rodolfo Facchini, Santa Marta, and Antonio Manes, "Acoustic Sensor Network for Vehicle Traffic Monitoring," *IARIA*, pp. 1–6, 2012.

[11] "http://audacity.sourceforge.net/," .

[12] Jong-seok Lee and Cheol Hoon Park, "Adaptive Decision Fusion for Audio-Visual Speech Recognition," 2008.

[13] T Lewis and D Powers, "Sensor fusion weighting measures in audio-visual speech recognition," in *Proceedings of the Conference on Australasian Computer Science*, 2004, pp. 305–314.

[14] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland, "Multimodal Person Recognition using Unconstrained Audio and Video," in *International Conference on Audio- and Video-Based Person Authentication*, 1999, pp. 176–181.

[15] Belur V Dasarathy, "Sensor Fusion Potential Exploitation Innovative Architectures and Illustrative Applications," *Proceedings of IEEE*, vol. 85, no. 1, 1997.