Applied Acoustics 138 (2018) 80–91

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

# Smartphone based traffic state detection using acoustic analysis and crowdsourcing

Dinesh Vij*, Naveen Aggarwal

*UIET, Panjab University, Chandigarh, India*

ARTICLE INFO

*Keywords:*
Traffic state detection
Smartphones
Microphone sensor
Wavelet packets
Crowdsourcing

ABSTRACT

Typically an Intelligent Transportation System (ITS) employs various types of infrastructure-based technologies into vehicles and roadways for monitoring traffic. But these solutions have high installation, maintenance, and operational costs. Further, most of these solutions are based on the assumptions of lane-based organized and homogeneous traffic, due to which these are ineffective in less organized traffic conditions which are common in developing countries. In this paper, we have proposed a cost-effective approach to infer the traffic state of the road by analyzing the cumulative acoustic signal collected from the microphone sensor of the user's smartphone. To capture the distinctive characteristics of various traffic scenes, we explored two different types of features: Mel Frequency Cepstral Coefficients (MFCCs) and Wavelet Packet Transform (WPT). Based on the understanding of acoustic signals of different traffic scenes, various parameters of these features such as window size and MFCC dimensions are tuned for better detection accuracy and robustness. To validate the approach, field experiments were conducted in varied conditions on the roads of City X. Experimental results revealed that for binary traffic scene classification *('busy-street'* vs. *'quiet-street')*, MFCC features are sufficient to get an overall accuracy of 100%. However, for *'congestion'* vs. *'medium-flow'* vs. *'free-flow'* traffic scene classification, MFCC features yield a bit lower accuracy of 77.64%. In this scenario, it was observed that WPT features can be used to reduce the false positive rate, thereby providing an absolute gain of 11.38% in the classification accuracy over the MFCC baseline. Also, it has been observed that by crowdsourcing the traffic state information from multiple users' smartphones, an effective accuracy improvement can be achieved for each traffic scene.

## 1. Introduction

Traffic congestion occurs when the traffic density on a particular road segment exceeds the available road capacity. Sometimes this situation is recurring in nature during office arriving and leaving hours (rush hours). At other times, it may be due to specific incidents such as accidents, road maintenance works, or weather conditions. As a consequence, it leads to many adverse effects such as time wastage, high fuel consumption, increased air and noise pollution, health consequences, wear and tear of vehicles, blocked emergency vehicles, stressed and frustrated commuters, and road rage. This inimical problem can be tackled by building new roads or widening the existing ones, or with better management of existing road infrastructure using Intelligent Transportation System (ITS) technologies [1]. The approach of increasing the road capacity is constrained by the availability of space. Therefore in the recent years, the emphasis is on effective management of existing road infrastructure. Though the infrastructure-based solutions employed in vehicles and roadways are widely used for monitoring traffic, these have high installation, maintenance, and operational costs. Furthermore, any technology based on the assumptions of lane-based, organized, and homogeneous traffic is unsuitable and inefficient for use in developing countries.

Several solutions have been proposed by the researchers that use various smartphone sensors such as GPS, Wi-Fi, and GSM for detecting the traffic states [2–7]. Recently, few researchers have proposed to use energy efficient sensors such as accelerometer and barometer for traffic state detection [8,9], because continuous sensing using location based sensors drains the battery of smartphones and the users might refrain from turning these sensors on. However, these techniques are based on the detection of movement and speed of the commuter, and require a significant amount of time for detecting the final traffic state [10]. One way to address this problem is to use acoustic analysis by collecting acoustic data from multiple microphones installed at roadsides in a particular fashion [11–13]. Acoustic analysis, also known as acoustic scene and event detection, has many advantages over other approaches of traffic state detection. Audio sensors are relatively less expensive and

* Corresponding author.
*E-mail addresses:* vijdinesh@gmail.com (D. Vij), navagg@gmail.com (N. Aggarwal).

https://doi.org/10.1016/j.apacoust.2018.03.029
Received 31 January 2017; Received in revised form 2 March 2018; Accepted 27 March 2018

consume less energy, i.e., ∼101 mW [14]. These are not affected by visual occlusions, orientation, and lighting conditions. These have low wear and tear and have the ability to work 24/7. This concept works even for non-lane driven traffic conditions. Moreover, it does not rely on the movement of the smartphone user for predicting different traffic states. In the audio signals, various types of events are present, such as engine start, engine stop, horn, music, and speech. These are known as acoustic events. These events usually occur within an environment type, known as acoustic scene, such as a car, bus, train, or street. Few researchers have worked on the detection of one particular type of traffic noise (acoustic event), such as honk signal, for inferring the traffic state (acoustic scene) of the road [15,16]. But, the use of one particular type of traffic noise alone may not provide good estimates of traffic. In 2012, Tyagi et al. [17] proposed to use cumulative road acoustics instead of one particular type of traffic noise. However, they used roadside-installed microphones for collecting acoustic data, which are rather expensive to install and maintain.

In this paper, we have proposed a cost-effective approach to determine the traffic state of the road using smartphones' microphone sensor. It is based on crowdsourced acoustic signal collected from multiple users' smartphones instead of microphones sensors installed at roadsides. The overall traffic state detection problem has been divided into two modules, as shown in Fig. 1. The module 1 remains active all the time, and module 2 is used only in case of detection of traffic (*'busy-street'*) on a particular road segment. For example, during night, there is less traffic on roads, so only module 1 would be used. It helps to reduce the computational complexity and avoids unnecessary battery drainage of smartphones.

In the first module, two broad traffic scenes have been considered: *'busy-street'* (traffic) and *'quiet-street'* (no traffic). *'Busy-street'* traffic scene is characterized by the sound signature of various traffic noises such as different vehicles passing by, tire noise, etc., whereas *'quiet-street'* traffic scene consists mainly of environmental noise and some traffic noise. Upon detection of *'busy-street,'* the traffic state detection problem is then sub-divided into three parts: *'congestion,' 'medium-flow,'* and *'free-flow,'* based on the traffic density of the road segment. The final detected class label along with the location of the mobile client (obtained by triggered GPS sensing parsimoniously) is then sent to a back-end server. The traffic state information from multiple users' smartphones, who are present at the same location, is then combined using crowdsourcing that can further be used by the traffic planners for high-level decision-making. For such crowdsourcing to be feasible on a large scale, a low-cost and energy-efficient sensor such as a microphone is required, which is ubiquitously available in every smartphone.

The field of acoustic scene and event detection is derived from the field of Automatic Speech Recognition (ASR). The features which have been widely used in speech recognition community, such as Mel Frequency Cepstral Coefficients (MFCCs), have also been used by researchers in the field of acoustic scene and event detection [17–19]. In this work, we have evaluated the applicability of MFCC features for detection of various traffic scenes. Various parameters of these features such as window size and MFCC dimensions have also been tuned for better detection accuracy and robustness. The large input audio signal is segmented into smaller manageable chunks for further processing. This process is known as windowing. The spectral characteristics of cumulative traffic scenes do not change significantly over short spans of time. Therefore, we analyzed the effect of larger window sizes on the accuracy of traffic scene detection. Also, traffic noise usually consists of low frequencies rather than high frequencies, and higher MFCC coefficients represent fast changes in the signal. So, the minimum number of MFCC coefficients required for effective detection of various traffic scenes has also been analyzed. Experimental results revealed that MFCC features based on the characteristics of speech spectrum have limited power in recognizing different traffic scenes and are not able to effectively distinguish between *'congestion'* and *'medium-flow'* classes. *'Medium-flow'* traffic scene is characterized by some concentration of spectral energy in low-frequency ranges and several high-frequency peaks over time, whereas *'congestion'* is characterized by a high concentration of spectral energy in low-frequency ranges due to engine idling. Though, few events such as high-frequency honk signals, are common to both *'congestion'* and *'medium-flow'* traffic scenes, and often lead to confusion between the two scenes. Further, MFCC features are extracted by computing Fourier transform of a signal. However, it has been shown that the Wavelet transform offers more flexibility to adapt its time–frequency resolution to the time–frequency characteristics of the non-stationary signals such as acoustic signals in a traffic scenario [20]. Wavelet transform has a property of multi-resolution, due to which it could be used to control time and frequency resolution in each spectral band effectively. The wavelet transform is mainly of two types: Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). The CWT is an analytical transform, and the DWT is a discrete implementation of the CWT. As an alternative, in Wavelet Packet Transform (WPT), both the detail and approximation coefficients are decomposed to create a full binary tree. This offers more flexibility to adapt the time frequency resolution of the analysis in comparison to the DWT [21]. So, we explored the use of WPT features in combination with MFCCs for a comprehensive description of the spectrum, which helped in improving the traffic state detection accuracy. The main contributions of this paper are summarized as follows:

- Use of multiple smartphones to detect the traffic states, thereby obviating the need for fixed infrastructure exhaustively at every place.
- Use of multi-level classification for traffic state detection, which helps to reduce the computational complexity and avoids unnecessary battery drainage of smartphones.
- Selection of an appropriate window size for different traffic scenes, based on the understanding of acoustic signals of these scenes.
- Selection of an appropriate feature set for different categories of traffic scenes so as to reduce the computational complexity and improve the detection capability of the predicting model. The use of WPT features provided an absolute gain of 11.38% in the traffic scene classification accuracy over the MFCC baseline.
- Validation of the above approaches on IEEE DCASE standard dataset [22] and self-recorded dataset (acoustic data collected for the different traffic scenes on various roads of City X). We achieved an accuracy of 100% for *'busy-street'* vs. *'quiet-street'* traffic scene classification and ∼89% accuracy for *'congestion'* vs. *'medium-flow'* vs. *'free-flow'* traffic scene classification.
- Improvement in the accuracy of traffic state detection by crowdsourcing of the traffic state information from multiple users' smartphones. An overall accuracy of 100% has been achieved for each traffic scene using crowdsourcing.

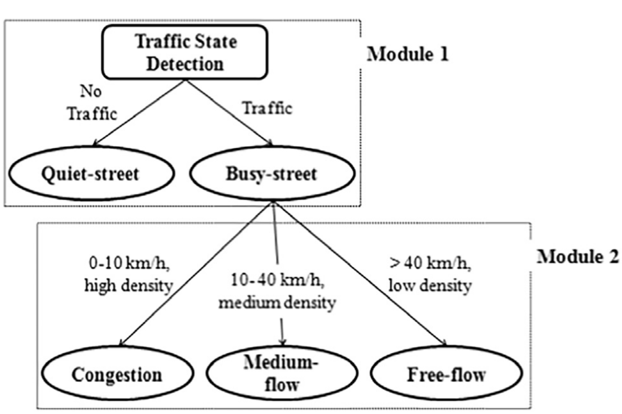


**Fig. 1.** Division of the detection system into modules.

The rest of the paper is organized as follows: Section 2 discusses the