TRAFFIC MONITORING WITH AD-HOC MICROPHONE ARRAY

Takuya Toyoda[1], *Nobutaka Ono*[2,3], *Shigeki Miyabe*[1], *Takeshi Yamada*[1], *Shoji Makino*[1]

[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
[3]The Graduate University for Advanced Studies (Sokendai)
toyoda@mmlab.cs.tsukuba.ac.jp, onono@nii.ac.jp,
{miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp

ABSTRACT

In this paper, we propose an easy and convenient method for traffic monitoring based on acoustic sensing with vehicle sound recorded by an ad-hoc microphone array. Since signals recorded by an ad-hoc microphone array are asynchronous, we perform channel synchronization by compensating for the difference between the start and the end of the recording and the sampling frequency mismatch. To monitor traffic, we estimate the number of the vehicles by employing the peak detection of the power envelopes, and classify the traffic lane from the difference between the propagation times of the microphones. We also demonstrate the effectiveness of our proposed method using the results of an experiment in which we estimated the number of vehicles and classified the lane in which the vehicles were traveling, according to F-measure.

Index Terms**—** Traffic monitoring, ad-hoc microphone array, synchronization, peak detection

1. INTRODUCTION

Monitoring traffic is important in relation to easing traffic congestion. Fixed-point observation techniques employ loop type sensor, ultrasonic sensor, infrared sensor, seismic sensor and movie cameras. However, these sensing systems suffer from high installation and maintenance costs. In addition, the accuracy of vehicle detection with these sensing systems depends heavily on the sensing condition. For example, with image processing, the quantity of data becomes enormous and accuracy degrades at night and in the bad weather. As a result, there has been a need for an easy method of monitoring traffic.

An effective monitoring approach should be inexpensive to install and maintain and have little dependence on the sensing environment. Several studies based on acoustic sensing have already attempted to meet these requirements [1, 2, 3, 4].

In this paper, for aiming easier and simpler sensing system, we present an acoustic traffic monitoring technique based on ad-hoc microphone array [5, 6] that combines independent recording devices for multichannel recording. The advantage of this approach is that its combination of small recording devices makes it easy to install. To synchronize the asynchronous recording channels [7, 8, 9], first we compensate for sampling mismatches by using single source activity [9]. Next, we analyze the power envelope of each channel and detect the peaks to estimate vehicle travel. By analyzing the time differences between the peaks of the channels, we obtain an estimation of the traffic lane in which the vehicle is traveling. Our experimental results confirm that we achieved highly accurate monitoring.

2. CHANNEL SYNCHRONIZATION

Since the proposed traffic monitoring includes analysis of time differences, synchronization of the asynchronous recording has to be corrected. In this section, we describe the flow of the synchronization using the two intervals of the recording of a single source activity proposed in [9].

Assume that sound pressures $\mathrm{x}_1(t)$ and $\mathrm{x}_2(t)$ on two microphones are sampled by different analog-to-digital converters (ADCs) as $x_1(n_1)$ and $x_2(n_2)$, where t and n_i (i = 1,2) denote continuous and discrete times series, respectively. Also assume that the sampling frequency of $x_1(n_1)$ is f_s, and that of $x_2(n_2)$ is $(1+\epsilon)f_s$ with a non-dimensional number ϵ. This paper also assumes that the ADCs have common nominal sampling frequencies and $|\epsilon| \ll 1$. Then, the relations between $x_i(n_i)$ and $\mathrm{x}_i(t)$ (i = 1,2) is given by

$$x_1(n_1) = \mathrm{x}_1\left(\frac{n_1}{f_s}\right), \tag{1}$$

$$x_2(n_2) = \mathrm{x}_2\left(\frac{n_2}{(1+\epsilon)f_s} + \Delta T_{21}\right), \tag{2}$$

where $\triangle T_{21}$ is the time at which the sampling of $x_2(n_2)$ starts. Here, the sample number that refers to the same time t of channel i (i = 1,2) is given by

$$n_1 = tf_s, \tag{3}$$

$$n_2 = (1+\epsilon)(t - \Delta T_{21})f_s \tag{4}$$

Then, n_2 is expressed with n_1 as below,

$$n_2 = (1+\epsilon)(n_1 - D_{21}), \tag{5}$$

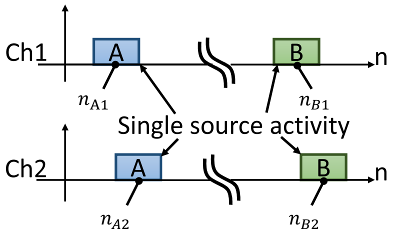

Fig. 1. Two pairs of single source activities

where $D_{21} = \Delta T_{21} f_s$ stands for the discrete time of the first channel when the recording of the second channel starts. Then, we show that both the sampling frequency mismatch ϵ and the recording start offset D_{21} are identifiable if two pairs of times $\{ n_{A1}, n_{A2} \}$ and $\{ n_{B1}, n_{B2} \}$ correspond to the same analogue time. These variables have to satisfy the following conditions.

$$n_{A2} = (1+\epsilon)(n_{A1} - D_{21}), \tag{6}$$

$$n_{B2} = (1+\epsilon)(n_{B1} - D_{21}). \tag{7}$$

The conditions identify ϵ and D_{21} as

$$\epsilon = \frac{n_{B2} - n_{A2}}{n_{B1} - n_{A1}} - 1, \tag{8}$$

$$D_{21} = \frac{n_{A1}n_{B2} - n_{A2}n_{B1}}{n_{B2} - n_{A2}}. \tag{9}$$

Thus, it is clear that we can obtain estimations of ϵ and D_{21} based on n_{Ai} and n_{Bi}. We can obtain estimations of ϵ and D_{21}. We estimate n_{Ai} and n_{Bi} by utilizing the relationship between two pairs of single source activities as shown in Fig. 1. We can estimate n_{Ai} and n_{Bi} by calculating the cross-correlation functions between each single source activities. According to the relations between n_1 and n_2, ϵ and D_{21}, we can compensate for the difference between the start and end of the recording.

Regarding compensation for the sampling frequency mismatch, we apply the linear phase $\exp(-jf\epsilon)$ to the signal of the second channel in the short-time Fourier transform (STFT) domain with the sampling frequency mismatch ϵ, where $j = \sqrt{-1}$ and f denote the frequency, and we obtain the signal where the sampling frequency mismatch has been compensated by inverse STFT analysis.

3. ACOUSTIC EVENT COUNTING

This section establishes a method for estimating the number of the vehicles according to peak detection of the power envelopes.

3.1. Approach

Since the captured sound power is expected to reach the maximum when vehicle is passing in front of the microphone, we try to count the number of vehicles by detecting the peaks of captured sound. However, due to the influence of background noise and fluctuations of observations, many spurious peaks appear in the power of raw data. Therefore, it is necessary to suppress the background noise and apply smoothing to the power before peak detection.

3.2. Noise suppression by Wiener filter

To suppress the noise, a standard way is to focus on the difference of the power spectra of target signal and noise. Generally, the sound of vehicles and background noise have different spectra. In this case, Wiener filter is one of the standard way, which is the optimum linear time-invariant filter in the sense of minimum mean square error. We manually find the frames where the vehicle is traveling and not traveling, and let $S_V(\omega)$ and $S_N(\omega)$ be the average vehicle sound and background noise power spectra, respectively. Here $\omega = 1, \cdots, \Omega$ denotes the discrete frequency index. We designed the following Wiener filter $W(\omega)$, assuming the stationarity of the power of the vehicle sound and background noise, to enhance the bands with a high SNR.

$$W(\omega) = \frac{S_V(\omega)}{S_V(\omega) + S_N(\omega)}. \tag{10}$$

Then, let $X_i(\omega, k)$ be the observed signal at the i-th microphone and an angular frequency ω in the k-th frame, and the noise suppressed signal $\hat{X}_i(\omega, k)$ is given as

$$\hat{X}_i(\omega, k) = W(\omega) X_i(\omega, k). \tag{11}$$

3.3. Gaussian smoothing and peak detection of temporal power envelope

To count acoustic events, we consider signal power time series $Y_i(k)$ given by

$$Y_i(k) = \frac{1}{\Omega} \sum_{\omega=1}^{\Omega} \hat{X}_i(\omega, k) \hat{X}_i^*(\omega, k), \tag{12}$$

where $\{\cdot\}^*$ denotes a complex conjugation. Even though a straightforward way is to count acoustic events directly from $Y_i(k)$, minute frequent fluctuations appear over $Y_i(k)$, which disturb the count. Hence, to reduce such frequent fluctuations, we calculate the power envelopes by smoothing. We perform smoothing with a Gaussian-window-shaped filter $g(m)$ given by

$$g(m) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{m^2}{2\sigma^2}\right), \tag{13}$$

where m is the time frame, and σ is the standard deviation of the Gaussian window. The signal smoothed by the Gaussian-window-shaped filter $\hat{Y}_i(k)$ is given by

$$\hat{Y}_i(k) = \sum_{m=\frac{-G}{2}}^{\frac{G}{2}} Y_i(k-m) g(m), \tag{14}$$

where G is the width of the Gaussian window. Even with such the smoothing, there still remain many peaks with small values. To ignore such small peaks, we discard the peaks with powers smaller than a threshold. If $\hat{Y}_i(k)$ is the maximum value and is more than h, let the time of peak k be $p_i(e)$,

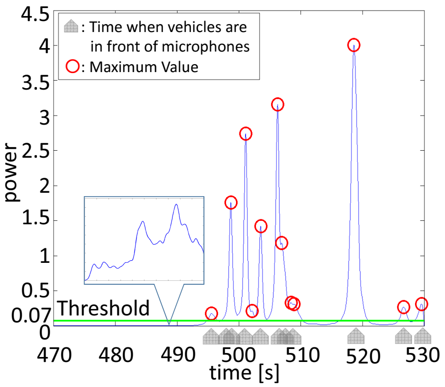

Fig. 2. Detected peaks with larger power than threshold

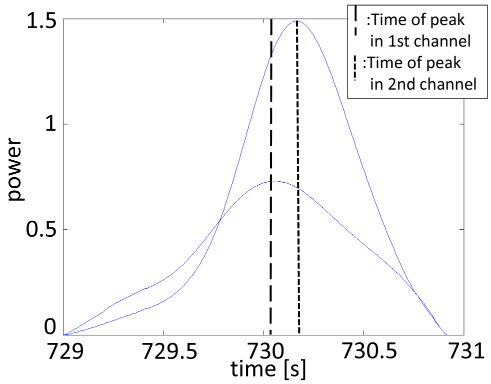

Fig. 3. Difference of peaks between channels

where h stands for the threshold, $p_i(e)$ for $e = 0, \cdots, E-1$ stand for the time of the detected peaks at i-th microphone, and E is the total number of detected peaks. The peak detection result is shown in Fig. 2. We confirmed that we can successfully detect peaks only when the vehicles are in front of the microphones.

4. TRAFFIC LANE ESTIMATION BASED ON TIME DIFFERENCE OF POWER PEAKS

This section establishes a method for estimating in which traffic lane the vehicle is traveling. We use the signals recorded by microphones that are installed in approximately same position of either side of two-lane roads. Since the distance to each microphone varies depending on the traffic lanes, there are differences in the propagation times between the two microphones according to the traffic lane in which the vehicles is traveling as shown in Fig. 3. So, we classify the traffic lanes in which the vehicle is traveling by using the differences in propagation time. As a first step, we associate two detected peaks at each microphone derived from the same vehicle. With $p_i(e)$ introduced in Sect **3.3**, if $|p_1(e) - p_2(e)|$ is less than d, let $p_1(e)$ be $\hat{p}_1(e)$, and $p_2(e)$ be $\hat{p}_2(e)$, where $\hat{p}_i(e)$ stands for the times of associated peaks. Range d is an arbitrary constant, which is determined in consideration of the speed of the traveling vehicle, the width of the road, the distance between microphones, the sampling frequency, and the frame length of STFT. When there are many peaks meeting those conditions, we choose peaks whose time difference is minimum. Then, we classify the traffic lane in which the vehicle is traveling with the following procedure. If $(\hat{p}_1(e) - \hat{p}_2(e))$ is less than 0, we classify the lane as the left traffic lane. If $(\hat{p}_1(e) - \hat{p}_2(e))$ is more than 0, we classify the lane as the right traffic lane.

5. PRELIMINARY EXPERIMENT IN REAL ROAD ENVIRONMENT

To confirm the effectiveness of our proposed acoustic sensing technique, we evaluated its accuracy in acoustic event counting and traffic lane estimation.

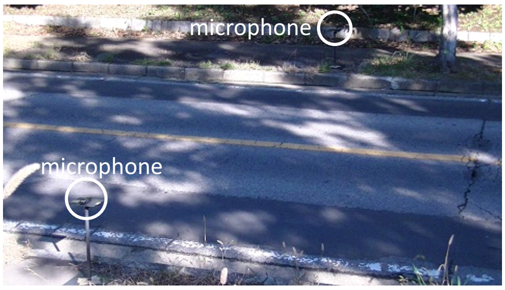

Fig. 4. A photo of recording setup

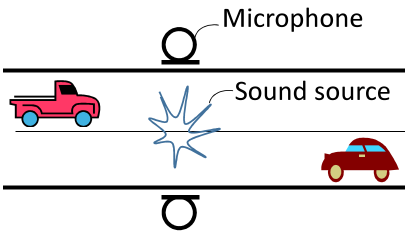

Fig. 5. Generating a single source activity for synchronization

5.1. Experimental conditions

We conducted experiment comparing an estimation obtained with our proposed acoustic sensing technique and the correct road traffic information regarding the number of traveling vehicles and the traffic lane in which they were traveling. We recorded the vehicle sounds of vehicles with ad-hoc microphone pairs and a movie to obtain the correct road traffic information on two-lane roads near Tsukuba university as shown in Fig. 4. In our experiments, we clapped hands at the center of the road shown in Fig. 5 and utilized these sounds as single source activities. We made a ground-truth data based on the movie and temporal waveform of sound power. After omitting intervals containing the sounds of motorcycles or bicycles, we estimated the number of vehicles and their traffic lanes from the recorded sound by every 60 seconds. For evaluation, we utilized one second interval as a tolerance, which means that the detection within ± 0.5 second from the ground-truth was counted as true detection. We evaluated our proposed acoustic sensing technique using F-measure obtained

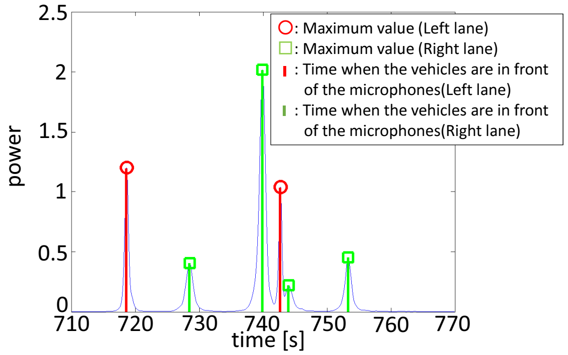

Fig. 6. Experimental result of traffic lane estimation

Table 1. Experimental conditions

Road width	7.1 m
Distance between microphones	9.5 m
Recording time	660 s
Sampling frequency	48 kHz
Frame lengths of STFT	2048 samples
Frame shift widths of STFT	512 samples
Standard deviation σ of Gaussian window	14
Width of Gaussian window	6σ+1
Threshold h	0.07
Range d of peak pairing	38 frames
Recording devices	SANYO ICR-PS603RM
Video camera	JVC GZ-HM670

from our comparison of the estimation and the correct traffic information provided by the reference movie. We experimentally determined the parameters σ and h by using the whole data. Therefore, this is a closed evaluation. An open evaluation is included in future work. The other experimental conditions are shown in Table 1. Because the time differences of power envelopes were used, the lane estimation should be robust even if the small mismatch of synchronization remains.

5.2. Calculation of F-measure

F-measure is used as a measure for a comprehensive evaluation of accuracy and completeness.

$$\text{precision} = \frac{N_c}{N_e} \tag{15}$$

$$\text{recall} = \frac{N_c}{N_r} \tag{16}$$

$$\text{F-measure} = \frac{\text{precision} \cdot \text{recall}}{\frac{1}{2}(\text{precision} + \text{recall})} \tag{17}$$

where N_c is the correctly estimated number, N_e is the total estimated number, and N_r is the total true number. F-measure has a value range of 0 to 1, and a higher value means better accuracy.

Table 2. Number of vehicles and detected vehicles(T = True, F = False)

		Traveled vehicles	
		T	F
Detected vehicles	T	61	8
	F	3	

Table 3. Number of vehicles and detected vehicles in each traffic lane

Traffic lane	Right			Left			
	Traveled vehicles				Traveled vehicles		
		T	F			T	F
Detected vehicles	T	22	3	Detected vehicles	T	35	4
	F	4			F	3	

Table 4. Calculated F-measure

	F-measure
Acoustic event counting	0.9173
Traffic lane estimation	0.8906

5.3. Experimental result and consideration

Fig. 6 shows estimation results obtained with the proposed acoustic sensing technique at certain time intervals. The result shows that both acoustic event counting and traffic lane estimation are performed well by our approach. As regards the evaluation over all the recording intervals, Tables 2 and 3 show the total true number of vehicles, the number of estimated vehicles and the number of correctly estimated vehicles counted for the two criteria. Table 4 shows the F-measure for acoustic event counting and traffic lane estimation. These results confirmed the appropriateness of the estimation obtained with our proposed acoustic sensing technique, based on peak detection and the time difference between signal power peaks, because high F-measures were obtained.

6. CONCLUSION

In this paper, we proposed a new traffic monitoring framework based on the sounds of vehicles traveling recorded with an ad-hoc microphone array. We performed channel synchronization based on single source activity. With our approach, we estimated the number of vehicles, and classified the traffic lane in which the vehicles were traveling. Since the F-measure of the experimental results achieved high values, we confirmed the effectiveness of our proposed method.

7. ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Research (B) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 25280069).

References

[1] N. Shimada, A. Itai and H. Yasukawa, "A study on an approaching vehicle detection using a linear microphone array-based acoustic sensing," *IEICE Technical Report SIS2009-71*, pp. 125–128, 2010. (in Japanese)

[2] Y. Nooralahiyan, M. Dougherty, D. Mckeown and R. Kirby, "A field trial of acoustic signature analysis for vehicle classification," *Transpn Res-C*, vol. 5, no. 3/4, pp. 165–177, 1997.

[3] M. Sobreira, A. Rodriguez and J. Alba, "Automatic classification of traffic noise," *Proc. IOA*, pp. 6221–6226, 2008.

[4] N. Evans and D. Chesmore, "Automated acoustic identification of vehicles," *Proc. IOA*, pp. 238–245, 2008.

[5] E. Robledo-Arnuncio, T. S. Wada and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Proc. WASPAA*, pp. 34–37, Oct. 2007.

[6] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, 2008.

[7] S. Markovich-Golan, S. Gannot and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," *Proc. IWAENC*, 2012.

[8] S. Miyabe, N. Ono and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," *Proc. ICASSP*, pp. 674–678, 2013.

[9] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity, " *Proc. APSIPA*, pp. 1–6, 2013.