Discrimination of Speech Activity and Impact Noise Using an Accelerometer and a Microphone in a Car Environment

Seon Man Kim[1], Hong Kook Kim[1], Sung Joo Lee[2], and Yun Keun Lee[2]

[1] School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea
{kobem30002,hongkook}@gist.ac.kr
[2] Speech/Language Information Research Center
Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea
{lee1862,yklee}@etri.re.kr

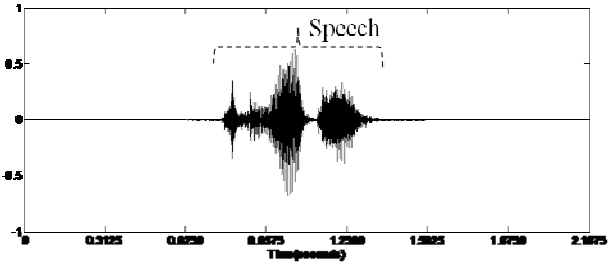
Abstract. In this paper, we propose an algorithm to discriminate speech from vehicle body impact noise in a car. Depending on road conditions such as the presence of large bumps or unpaved stretches, impact noises from the car body may interfere with the detection of voice commands for a speech-enabled service in the car, which results in degraded service performance. The proposed algorithm classifies each analysis frame of the input signal recorded by a microphone into four different categories such as speech, impact noise, background noise, and mixed speech and impact noise. The classification is based on the likelihood ratio test (LRT) using statistical models constructed by combining signals obtained from the microphone with those from an accelerometer. In other words, the different characteristics detected by both acoustical and mechanical sensing enable better discrimination of voice commands from noise emanating from the vehicle body. The performance of the proposed algorithm is evaluated using a corpus of speech recordings in a car moving at an average velocity of 30-50 km/h with impact noise at various signal-to-noise ratios (SNRs) from -3 to 1 dB, where the SNR is defined as the ratio of the power of speech signals to that of impact noise. It is shown from the experiments that the proposed algorithm achieves a discrimination accuracy of 85%.

Keywords: Speech enabled service in a car, car impact noise, voice activity detection, accelerometer.

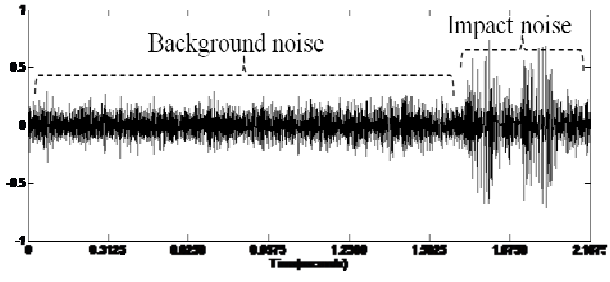
1 Introduction

Interest in speech interfaces for controlling electronic products has grown rapidly because of safety and convenience concerns. In a vehicle, it is particularly essential for the driver to use a speech interface system to control electronic devices, e.g., car navigation or telematics systems. However, the quality of the speech signal in car environments is deteriorated by the numerous noise sources such as the car engine, fan, audio system, wind, road, and conversation among passengers [1-3].

T.-h. Kim et al. (Eds.): FGCN 2011, Part II, CCIS 266, pp. 104–113, 2011.



(a)



(b)

Fig. 1. Examples of speech, background noise, and impact noise recordings: (a) male speech, (b) noise when driving at 30-50 km/h

Road noise is resulted from the movement of a vehicle's tires over the road surface, and it is the major source of stationary background noise exposure. In particular, the tires' contact with a speed bump or barrier on a road induces vehicles to vibrate, which brings about impact noises [4-6]. Furthermore, front and rear tires of vehicles contribute to two successive impact noises, having durations and waveforms that are similar to speech, as shown in Fig. 1. Even though impact noises degrade the performance of speech interface systems such as hand-free communication systems and speech recognition systems, few studies have been conducted which deal with impact noises. Therefore, the development of a method of discriminating speech signal from impact noise is required for the realization of successful speech interface systems in car environments [7-10].

Therefore, this paper proposes an algorithm for discriminating impact noise and driver's speech signal in a car environment. This discrimination is accomplished by the combination of three decision rules pertaining to non-background noise, impact noise, and speech. Recently proposed statistical model-based decision rules have demonstrated good performance by employing the likelihood ratio test (LRT) with the complex Gaussian distribution [11]. In our experience, however, the statistical model-based detectors were incapable of discriminating impact noises from speech because two signals were very similar each other, as shown in Fig. 1. Thus, it is necessary to obtain information on impact noise in an alternative way, which should not be significantly affected by speech signal. Fortunately, the tires' contact with a speed bump or

barrier on a road induces impact noise that is further transmitted as vehicle shock vibration over the car body [4]. This transmitted vibration can be easily measured by an accelerometer. Thus, we can propose an impact noise activity detector comprising the signals from an accelerometer instead of those from a microphone.

The remainder of this paper is organized as follows. Following this introduction, we review a statistical model-based decision rule in Section 2. In Section 3, we propose a technique for discriminating impact noises from speech. In particular, the approach to utilizing an accelerometer in detecting impact noise activity is proposed. In Section 4, we evaluate the discrimination performance of the proposed method. Finally, we summarize our findings in Section 5.

2 Statistical Model-Based Target Signal Activity Detection

Target signal activity detection can be interpreted as a binary hypothesis test. Let $X_k(\ell)$ be the spectral component of a microphone signal, where $k(=1,2,\cdots,K)$ is a frequency bin index and $\ell(=1,2,\cdots)$ is a frame index. Also, let $T_k(\ell)$ and $D_k(\ell)$ denote the target and non-target spectral component, respectively. Then, two hypotheses can be described as $H_{T,0}: X_k(\ell) = D_k(\ell)$ and $H_{T,1}: X_k(\ell) = D_k(\ell) + T_k(\ell)$. Assuming that $T_k(\ell)$ and $D_k(\ell)$ follow zero-mean complex Gaussian distributions, the likelihood ratio on $H_{T,0}$ and $H_{T,1}$ under the observation $X_k(\ell)$, $\Lambda_k(X_k(\ell))$, is given by,

$$\Lambda_k(X_k(\ell)) = \frac{1}{1+\psi_{T,k}(\ell)} \exp\left(\frac{\varphi_{T,k}(\ell)\psi_{T,k}(\ell)}{1+\psi_{T,k}(\ell)} \right) \tag{1}$$

where $\psi_{T,k}(\ell) = \lambda_{T,k}(\ell)/\lambda_{D,k}(\ell)$ and $\varphi_{T,k}(\ell) = |X_k(\ell)|^2/\lambda_{D,k}(\ell)$. Here, $\psi_{T,k}(\ell)$ and $\varphi_{T,k}(\ell)$ indicate the *a priori* and *a posteriori* signal-to-noise ratio (SNR), respectively. In addition, $\lambda_{T,k}(\ell)$ and $\lambda_{D,k}(\ell)$ indicate target and non-target spectral variance, respectively.

Then, a target activity decision rule is established from the average value of the log likelihood ratio for individual frequency bin as,

$$\log \Lambda_T(\ell) = \frac{1}{K}\sum_{k=0}^{K-1} \log \Lambda_k(\ell) \quad \begin{matrix} \geq \ \eta_T : \ H_{T,1} \\ < \ \eta_T : \ H_{T,0} \end{matrix} \tag{2}$$

where η_T is a pre-determined decision threshold. Consequently, the decision rule strongly depends on the reliable estimate of $\lambda_{T,k}(\ell)/\lambda_{D,k}(\ell)$ and $\psi_{T,k}(\ell)/\varphi_{T,k}(\ell)$. In other words, the statistical model-based target signal activity decision rule requires target and non-target signal spectral variance estimation. Therefore, the decision procedure is rewritten as

$$\{H_{T,0} \ \text{ or } \ H_{T,1}\} = \Re_T(\psi_{T,k}(\ell), \varphi_{T,k}(\ell), \zeta_D, \eta_T)_{\forall k} \tag{3}$$

where $\Re_T(\cdot)$ is a statistical model-based decision function and $\forall k$ means that all frequency bins are used to decide the target signal activity.

3 Proposed Impact Noise and Speech Discrimination Method

Let $N_k(\ell)$ denote the k-th spectral component of the ℓ-th frame of background car noise, distributed over the entire time interval, as shown in Fig. 1(b). In addition, let $V_k(\ell)$ and $S_k(\ell)$ be impact noise and speech of the k-th frequency bin and the ℓ-th frame, respectively. Then, depending on the presence or absence of $V_k(\ell)$ and $S_k(\ell)$, four hypotheses could be constructed as

$$\begin{aligned} &H_0 : X_k(\ell) = N_k(\ell) && H_1 : X_k(\ell) = N_k(\ell) + V_k(\ell) \\ &H_2 : X_k(\ell) = N_k(\ell) + S_k(\ell), && H_3 : X_k(\ell) = N_k(\ell) + S_k(\ell) + V_k(\ell). \end{aligned} \tag{4}$$

This paper aims to decide which hypothesis out of four hypotheses, i.e., H_0, H_1, H_2 and H_3, is true. Assuming that $N_k(\ell)$, $V_k(\ell)$ and $S_k(\ell)$ follow zero-mean complex Gaussian distributions, 16 likelihood ratios, i.e., Λ_{ij} (i=0,1,2,3; j=0,1,2,3), could be derived directly on H_0, H_1, H_2 and H_3 [12]. The approach using 16 likelihood ratios makes it difficult to tune the relevant parameters and optimize its performance. Instead, the 16 hypotheses can be reduced into six ones by properly combining three kinds of target signal activity hypothesis models. The first of them is the hypothesis of non-background noise, H_{VS}, which corresponds to any of impact noise, speech or both. On one hand, H_S and H_V are defined for only speech and impact noise activity, respectively. They are represented as

$$H_{VS,0} : X_k(\ell) = N_k(\ell), \quad H_{VS,1} : X_k(\ell) = N_k(\ell) + VS_k(\ell) \quad \text{with} \quad VS_k(\ell) = V_k(\ell) + S_k(\ell) \tag{5a}$$

$$H_{S,0} : X_k(\ell) = NV_k(\ell), \quad H_{S,1} : X_k(\ell) = NV_k(\ell) + S_k(\ell) \quad \text{with} \quad NV_k(\ell) = N_k(\ell) + V_k(\ell) \tag{5b}$$

$$H_{V,0} : X_k(\ell) = N_k(\ell), \quad H_{V,1} : X_k(\ell) = NS_k(\ell) + V_k(\ell) \quad \text{with} \quad NS_k(\ell) = N_k(\ell) + S_k(\ell). \tag{5c}$$

From Eqs. (5a), (5b) and (5c), we can further define four different hypotheses, including only background noise activity, H_0, speech activity, H_1, impact noise activity, H_2, and mixture of speech and impact noise activity, H_3. That is,

$$H_0 \leftarrow H_{VS,0} \;\&\; H_{V,0} \;\&\; H_{S,0}, \qquad H_1 \leftarrow H_{VS,1} \;\&\; H_{V,0} \;\&\; H_{S,1} \tag{6a}$$

$$H_2 \leftarrow H_{VS,1} \;\&\; H_{V,1} \;\&\; H_{S,0}, \qquad H_3 \leftarrow H_{VS,1} \;\&\; H_{V,1} \;\&\; H_{S,1}. \tag{6b}$$

Fig. 2 shows a block diagram of the proposed approach. First, the power spectral density (PSD) of the acoustic signal, $|X_k(\ell)|^2$, which is obtained from a microphone, is used to detect non-background noise and speech activity, i.e., $H_{VS,0}$ / $H_{VS,1}$. On the other hand, the PSD of the vibration signal, $|Y_k(\ell)|^2$, which is obtained from an accelerometer, is used to detect impact noise activity, $H_{V,0}$ / $H_{V,1}$. In addition, the

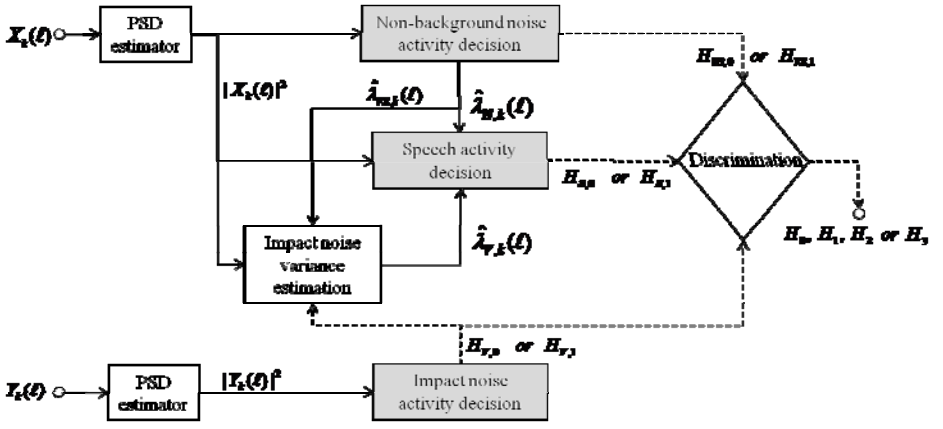


Fig. 2. Block diagram of the proposed discrimination approach between impact noise and speech

vibration signal-based impact noise activity decision, $H_{V,0}$ / $H_{V,1}$, and non-background noise and speech activity decision, $H_{VS,0}$ / $H_{VS,1}$, are also utilized together to estimate the spectral variance, $\lambda_{V,k}(\ell)$, of impact noise from a microphone, which is used to decide speech activity, i.e., $H_{S,0}$ / $H_{S,1}$. Finally, among H_0, H_1, H_2 and H_3, we decide which hypothesis is true by using the three decision results of $H_{VS,0}$ / $H_{VS,1}$, $H_{V,0}$ / $H_{V,1}$, and $H_{S,0}$ / $H_{S,1}$.

3.1 Statistical Model-Based Impact Noise and/or Speech Activity Detection

From the hypothesis on the presence, $H_{VS,1}$, or absence, $H_{VS,0}$, of non-background noise in Eq. (5a), we can detect non-background noise activity using a statistical model-based decision rule, as defined in Section 2. That is,

$$\{H_{VS,0} \quad \text{or} \quad H_{VS,1}\} = \Re_{VS}(\psi_{VS,k}(\ell), \varphi_{VS,k}(\ell), \eta_{VS})_{\forall k} \tag{7}$$

where $\Re_{VS}(\cdot)$ denotes a statistical model-based decision function for the non-background noise activity, $VS(\ell)$, and η_{VS} is a threshold.

3.2 Statistical Model-Based Speech Activity Detection

Contrary to Section 3.1, from the hypothesis on the presence, $H_{S,1}$, or absence, $H_{S,0}$, of speech in Eq. (5b), we can also detect speech activity using a statistical model-based decision rule, such as

$$\{H_{S,0} \quad \text{or} \quad H_{S,1}\} = \Re_{S}(\psi_{S,k}(\ell), \varphi_{S,k}(\ell), \eta_{S})_{\forall k} \tag{8}$$

where $\Re_S(\cdot)$ denotes a speech activity decision function for $S(\ell)$ and η_S is a threshold.

In Eq. (8), $\psi_{S,k}(\ell) = \lambda_{S,k}(\ell)/\lambda_{NV,k}(\ell)$ and $\varphi_{S,k}(\ell) = |X_k(\ell)|^2/\lambda_{NV,k}(\ell)$. In order to estimate $\psi_{S,k}(\ell)$ and $\varphi_{S,k}(\ell)$, the estimate of $\lambda_{NV,k}(\ell)$, $\hat{\lambda}_{NV,k}(\ell)$, is obtained by

$$\hat{\lambda}_{NV,k}(\ell) = \hat{\lambda}_{N,k}(\ell) + \hat{\lambda}_{V,k}(\ell) \tag{9}$$

where $\hat{\lambda}_{N,k}(\ell)$ and $\hat{\lambda}_{V,k}(\ell)$ are the PSD estimate of background noise and impact noise, respectively. Note that $\hat{\lambda}_{N,k}(\ell)$ has been already estimated from the non-background noise activity estimation procedure. In addition, we will give a detail explanation on estimating $\hat{\lambda}_{V,k}(\ell)$ in the next subsection. In order to estimate $\psi_{S,k}(\ell)$, $\hat{\lambda}_{S,k}(\ell)$ should be also estimated before, which is done by a spectral subtraction method between $\hat{\lambda}_{VS,k}(\ell)$ and $\hat{\lambda}_{V,k}(\ell)$ by the following equation of

$$\hat{\lambda}_{S,k}(\ell) = \max(\hat{\lambda}_{VS,k}(\ell) - \beta \cdot \hat{\lambda}_{V,k}(\ell), 0) \tag{10}$$

where β is a tuning parameter.

3.3 Impact Noise Activity Detection Using an Accelerometer

As mentioned in Section 1, we utilize an accelerometer for impact noise activity detection. Let $Y_k(\ell)$ be the spectral component of sensing signal by an accelerometer at the k-th frequency bin and ℓ-th frame. Also, let $N'_k(\ell)$ and $V'_k(\ell)$ denote the accelerometer noise (or impact noise) and impact vibration, respectively. Then, two hypotheses can be postulated as $H'_{V,0}: Y_k(\ell) = N'_k(\ell)$ and $H'_{V,1}: Y_k(\ell) = N'_k(\ell) + V'_k(\ell)$. Applying a statistical-based decision rule similar to Eq. (3), we can detect the impact vibration activity by

$$\{H'_{V,0} \text{ or } H'_{V,1}\} = \Re'_V(\psi'_{V,k}(\ell), \varphi'_{V,k}(\ell), \eta'_V)_{\forall k}, \tag{11}$$

where the distributions of $N'_k(\ell)$ and $V'_k(\ell)$ are zero-mean complex Gaussian, $\Re'_V(\cdot)$ denotes a decision function of an impact vibration activity, and η'_V is a threshold.

Figs. 3(a) and 3(b) show that the time interval of impact noise is similar to that of impact vibration, which is estimated by a log likelihood-based decision rule depicted in Fig. 3(c). Thus, upon this observation, we alternatively use the impact vibration activity interval instead of the impact sound activity interval in this paper.

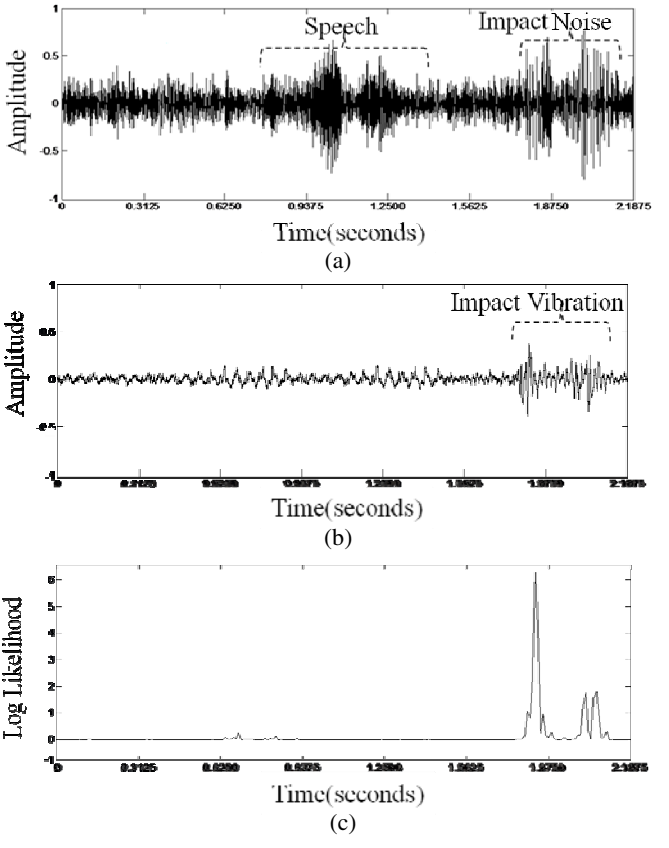


Fig. 3. Illustration of the relationship between impact noise and impact vibration; (a) the acoustic signal obtained from a microphone, (b) vibration signal obtained from an accelerometer and (c) log likelihood estimate of the impact vibration activity

Although the impact noise activity detection is approximately accomplished by detecting the impact vibration activity, the spectral variance estimate of the impact noise, $\hat{\lambda}_{V,k}(\ell)$, is required in Eqs. (9) and (10). In other words, we use a Wiener filter in order to estimate $\hat{\lambda}_{V,k}(\ell)$. Also, $\hat{\lambda}_{NS,k}(\ell)$ is estimated by a recursive procedure executed only when $H'_{V,0}$ is determined to be true. That is, we have

$$\hat{\lambda}_{NS,k}(\ell) = \zeta_{NS} \cdot \hat{\lambda}_{NS,k}(\ell-1) + (1-\zeta_{NS}) \cdot \hat{\lambda}_{VS,k}(\ell) \tag{12}$$

where ζ_{NS} is a smoothing parameter.

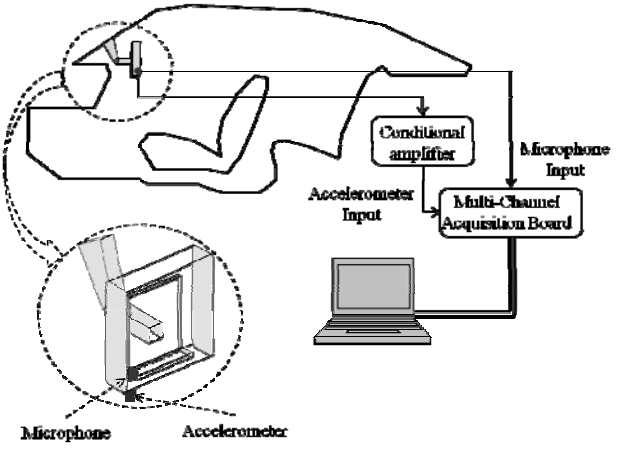


Fig. 4. Configuration of an accelerometer sensor and a microphone system in order to collect impact vibration and speech signals from a commercial navigation system

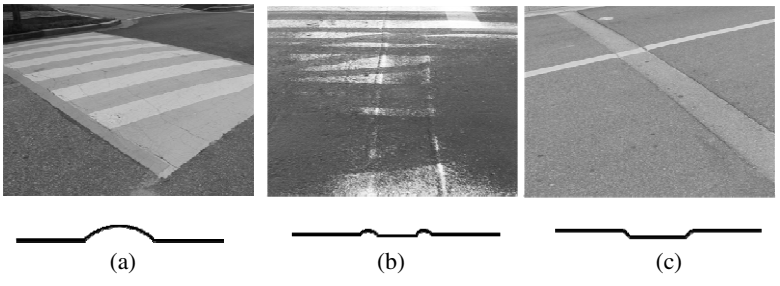


Fig. 5. Snapshot of the road conditions for impact noise; (a) a speed bump, (b) two successive convex hemispherical surfaces, and (c) a concave surface

4 Performance Evaluation

Fig. 4 shows a system employing a microphone and an accelerometer, which was constructed in a 2,000 cc class vehicle in order to collect test data for performance evaluation. In particular, a commercial navigation platform equipped with a low cost microphone was used to collect speech test data. Here, B&K type 4393 accelerometer sensor was used in order to measure the car impact vibration signal, where a B&K Type 2692 conditional amplifier was used to adjust the signal power. The desired speech signal was played from a speaker mounted below the headrest of the driver's seat under a clean condition. The background noise, impact noise, and vibration signal were recorded while driving the car at an average speed between 30 and 50 km/h subject to three different road conditions shown in Fig. 5.

Table 1. Confusion matrix between the manual decision and the decision by the proposed algorithm

Manual Decision	Decision by the Proposed Algorithm											
	Case I				*Case II*				*Case III*			
	H_0	H_1	H_2	H_3	H_0	H_1	H_2	H_3	H_0	H_1	H_2	H_3
H_0	91.4	1.8	6.6	0.1	92.7	2.4	4.7	0.3	95.6	1.6	2.8	0.05
H_1	10.3	85.2	3.5	0.9	7.3	89.6	0.3	2.8	9.9	76.2	1.7	12.2
H_2	6.1	0	93.9	1.0	3.3	0.9	88.5	7.2	-	-	-	-
H_3	-	-	-	-	-	-	-	-	0.3	4.2	10.4	85.2

Located on the dashboard, 10 speech utterances by 5 males and 5 females, 10 impact noise signals, and vibration signals were separately recorded at a sampling rate of 16 kHz. In order to simulate the driving conditions, we mixed speech signals with impact noise in three different ways. For the first case, impact noise occurred at the beginning of the speech interval, which was referred as *Case I*. For the second one, impact noise appeared at the end of the speech interval (*Case II*). Lastly, we added impact noise into the speech interval (*Case III*). Subsequently, 100 mixed utterances were prepared as a test database, where SNRs ranged from -3 dB to 1 dB.

The proposed method was applied once every frame whose length was 20 ms long. In order to obtain spectral components, we applied a 320-point discrete Fourier transform (DFT). Throughout the experiment, we set $\eta_{VS} = 0.5$ in Eq. (7), $\eta_S = 0.5$ in Eq. (8), $\beta = 0.5$ in Eq. (10), $\eta'_V = 0.15$ in Eq. (11), and $\zeta_{NS} = 0.9$ in Eq. (12).

The performance of the proposed method was measured as an accuracy ratio between the decision outputs from the proposed algorithm and those obtained manually. Table 1 summarizes the results in a confusion matrix form. As shown in the table, the proposed method exhibited more than 85% accuracy in *Case I* and *Case II*. The accuracy of *Case III* was lowered than those of *Case I* and *Case II*. This was mainly contributed by the confusion between H_1 and H_3 states.

5 Conclusion

In this paper, we proposed a technique to discriminate between impact noise and speech activity in a car environment. The proposed technique employed three statistical model-based decision rules, for non-background noise, impact noise and speech activities. In particular, an effective impact noise detector using an accelerometer was proposed. Then, each decision rule result was utilized to discriminate between an impact noise and speech activity. From the performance evaluation, the proposed algorithm had a discrimination accuracy of 85 %.

Acknowledgements. This work was supported in part by the Industrial Strategic technology development program, 10035252, 'Development of dialog-based spontaneous speech interface technology on mobile platform' funded by the Ministry of Knowledge Economy, Korea, and by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No.2011–0026201).

References

1. Wu, K.G., Chen, P.C.: Efficient speech enhancement using spectral subtraction for car hands-free applications. In: roceedings of International Conference on Consumer Electronics (ICCE), Las Vegas, NV, pp. 220–221 (2007)
2. Ahn, S., Ko, H.: Background noise reduction via dual-channel scheme for speech recognition in vehicular environment. IEEE Transactions on Consumer Electronics 51(1), 22–27 (2005)
3. Kim, S.M., Kim, H.K.: Hybrid probabilistic adaptation model controller for generalized sidelobe canceller-based target-directional speech enhancement. In: Proceedings of ICASSP, Prague, Czech Republic, pp. 2532–2535 (2011)
4. Lee, S.-K., Kim, H.-W., Na, E.-W.: Improvement of impact noise in a passenger car utilizing sound metric based on wavelet transform. Journal of Sound and Vibration 329(17), 3606–3619 (2010)
5. Lee, S.-K., Chae, H.-C.: The application of artificial neural networks to the characterization of interior noise booming in passenger cars. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 218(1), 33–42 (2004)
6. Wang, Y.S., Lee, C.-M., Kim, D.-G., Xu, Y.: Sound-quality prediction for nonstationary vehicle interior noise based on wavelet pre-processing neural network model. Journal of Sound and Vibration 299(4-5), 933–947 (2007)
7. Hu, J.S., Cheng, C.C., Liu, W.H., Yang, C.H.: A robust adaptive speech enhancement system for vehicular applications. IEEE Transactions on Consumer Electronics 52(3), 1069–1077 (2006)
8. Kim, S.M., Kim, H.K.: Probabilistic spectral gain modification applied to beamformer-based noise reduction in a car environment. IEEE Transactions on Consumer Electronics 57(2), 866–872 (2011)
9. Park, J.H., Kim, S.M., Yoon, J.S., Kim, H.K., Lee, S.J., Lee, Y.K.: SNR–based mask compensation for computational auditory scene analysis applied to speech recognition in a car environment. In: Proceedings of Interspeech, Makuhari, Japan, pp. 725–728 (2010)
10. Park, J.H., Shin, M.H., Kim, H.K.: Statistical model-based voice activity detection using spatial cues and log energy for dual-channel noisy speech recognition. CCIS, vol. 120, pp. 172–179 (2010)
11. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters 6(1), 1–3 (1999)
12. Lee, S.Y., Shin, J.W., Yun, H.S., Kim, N.S.: A statistical model based post-filtering algorithm for residual echo suppression. In: Proceedings of Interspeech, Antwerp, Belgium, pp. 858–861 (2007)