

A Study on CAN Signal Identification via Linear Regression Models

ByeongYeol Ahn¹, JaeHyeok Yang², SeoYeon Kim² and TaeGuen Kim^{3,*}

¹Department of Software Convergence, Soonchunhyang University, Asan, Korea
20184603@sch.ac.kr

²Department of Mobility Convergence Security, Soonchunhyang University, Asan, Korea
{seoyeon56, vmffs1t}@sch.ac.kr

^{3,*} Department of Information Security Soonchunhyang University, Asan, Korea
tg.kim@sch.ac.kr

Abstract

Identifying CAN data is crucial for ensuring the reliability and integrity of evidence in car accidents. The CAN signal identification process confirms the relevance of events from a forensic perspective, facilitates evidence collection, and ensures data integrity. In this paper, we present a method for analyzing CAN traffic data using a linear regression analysis model in scenarios where a CAN database is unavailable. Our approach aims to identify a signal representing vehicle speed within CAN traffic data. Building on this foundation, we anticipate the extra development of a model for automated interpretation of multiple signals that increase or decrease linearly.

Keywords:CAN protocol, CAN signal identification, Linear regression, Vehicle security

1 Introduction

CAN (Control Area Network) is an automotive network protocol used for sharing data between ECUs (Electronic Control Units). ECUs that are responsible for various vehicle functions, engage in cooperative control based on data exchanged through CAN communication. We present a method for analyzing the meaning of CAN signal data, using a linear regression model, without relying on a CAN database. After extracting sequences composed of CAN messages with the same ID, we divide each sequence data into data payloads. Subsequently, we employ a sliding window approach on sequences with varying elements to ascertain whether they represent speed data through a linear regression model.*

The 7th International Conference on Mobile Internet Security (MobiSec'23), Dec. 19-21, 2023, Okinawa, Japan, Article No.P-11

*Corresponding author Department of Information Security, Soonchunhyang University, Seoul, 31538, Republic of Korea, Email: tg.kim@sch.ac.kr

2 Our proposed method

In this paper, we introduce a new method that consists of two phases: Training phase, and Test phase. The training phase consists of three steps: Arbitrary signal sequence extraction, Linear model generation, R-squared score calculation. The comprehensive workflow of our proposed method is depicted in Figure 1. Initially, In the Arbitrary signal sequence extraction step, we initiate the process by isolating sequences of CAN messages that have identical ID. Then, the data payloads from these congruent ID messages are partitioned into uniform bit lengths. From these segmented data pieces, a novel sequence is generated based on the similar positions. We Extract sub-sequence is derived by implementing a sliding window technique on the variable sections of the newly formed sequence. In the Linear model generation step these processed features are modeled using both first-order and second-order linear regression models. In the R-squared score calculation step, we calculate all the R-squared scores between the data used to create the linear regression model and average R-squared scores.

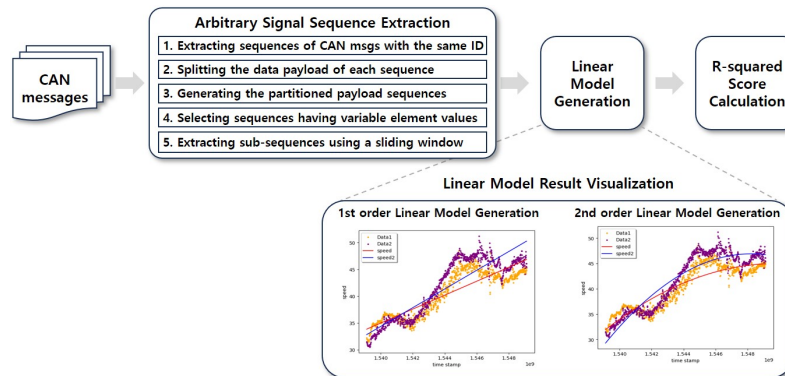


Figure 1: The processing flow of our proposed method

After modeling, we determined that the periodicity of the timestamps and the constant increase and decrease of the velocity values matched the characteristics of the velocity data. In the test phase, when an unknown sequence is given to be analyzed for signal type identification, the average of R-squared scores of the sequence is calculated by conducting three steps of the training phase. Then the difference between the average score of the unknown sequence and the one of velocity signal sequence. If the difference value is less than the pre-defined threshold, then the framework classifies the unknown sequence to the velocity sequence.

3 Evaluation

To assess the effectiveness of our proposed method, we measured the R-squared score values of the velocity sequence. The average of the R-squared scores was approximately 0.71 and 0.87 for the 1st-order linear regression model and the 2nd-order linear regression model, respectively. The 2nd-order model performed better than the 1st order model, and this may be due to the observation that acceleration value is not constant, because the acceleration depends on the driver's pressure applied to the accelerator. It is expected that it can be used as evidence for accident investigation such as identifying the cause of automobile accidents in the future.

References

- [1] Projectgus. (2023). *Hyundai-Kona- Can-Logs*. github.com/projectgus/hyundai-kona-ev-can-logs.
- [2] Socialledge. (October 11, 2017). *DBC*. http://socialledge.com/sjsu/index.php/DBC_Format.