International Journal of Automotive Technology (2024) 25:1025–1035
https://doi.org/10.1007/s12239-024-00078-8

NOISE, VIBRATION AND HARSHNESS

In-Vehicle Environment Noise Speech Enhancement Using Lightweight Wave-U-Net

Byung Ha Kang[1] · Hyun Jun Park[1] · Sung Hee Lee[2] · Yeon Kyu Choi[2] · Myoung Ok Lee[3] · Sung Won Han[1]

Received: 12 December 2022 / Revised: 12 December 2022 / Accepted: 20 March 2024 / Published online: 19 April 2024

Abstract
With the rapid advancement of AI technology, speech recognition has also advanced quickly. In recent years, speech-related technologies have been widely implemented in the automotive industry. However, in-vehicle environment noise inhibits the recognition rate, resulting in poor speech recognition performance. Numerous speech enhancement methods have been proposed to mitigate this performance degradation. Filter-based methodologies have been used to remove existing vehicle environment noise; however, they remove only limited noise. In addition, there is the constraint that there are limits to the size of models that can be mounted inside a vehicle. Therefore, making the model lighter while increasing speech quality in a vehicle environment is an essential factor. This study proposes a Wave-U-Net with a depthwise-separable convolution to overcome these limitations. We built various convolutional blocks using the Wave-U-Net model as a baseline to analyze the results, and we designed the network by adding squeeze-and-excitation network to improve performance without significantly increasing the parameters. The experimental results show how much noise is lost through spectrogram visualization, and that the proposed model improves performance in eliminating noise compared with conventional methods.

Keywords In-vehicle noise environment · Speech enhancement · Deep learning · U-Net · NVH · Time-domain · Convolutional neural network (CNN)

Abbreviations

ASR	Automatic speech recognition
CNN	Convolutional neural network
D	Discriminator
DEM	Discriminator–valuation mismatch
DNN	Deep neural network
G	Generator
GELU	Gaussian error linear units
GLU	Gated linear unit
MSE	Mean squared error
SE	Speech enhancement
STFT	Short-time Fourier transform
T–F	Time frequency

✉ Sung Won Han
swhan@korea.ac.kr

[1] School of Industrial and Management Engineering, Korea University, Seoul 02841, Korea

[2] Analytics Team, BRFrame Inc., Seoul 05836, Korea

[3] NVH Research Lab, Hyundai Motor Company, Hwaseong-si, Gyeonggi-Do 18270, Korea

1 Introduction

In recent years, with the rapid development of science and technology, voice interaction systems based on automatic speech recognition (ASR) have been widely used in many industries, including manufacturing, communication, and finance. Especially, in the automotive industry, a number of in-car applications use speech recognition, such as navigation systems, hands-free communications, and podcasts (Gellatly & Dingus, 1998). However, the degradation of ASR performance caused by ambient noise and reverberation, which corrupts the speech captured by microphones, remains one of the most significant factors that limit their performance. This may be the result of two types of noise: (1) external noise, such as wind noise at high speeds, and (2) internal noise, such as engine noise and air conditioning noise. In-vehicle environment noise often leads to a deterioration in speech recognition capabilities. Currently, the accuracy of speech recognition in vehicles exceeds 98% under ideal environmental conditions. However, as the environment changes, the vehicle speech recognition rate decreases, and speech recognition against a complex vehicle