Audio Engineering Society

Convention Express Paper 152

Presented at the 155th Convention

2023 October 25-27, New York, USA

Transformer-Based Virtual Engine Sound Generation Method

Jaeyoung Lee[1], Dooil Choi[1] and Jongin Jung[1]

[1] *Hyundai Mobis, 17-2 Mabuk-ro 240beon-gil, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea*

Correspondence should be addressed to Jaeyoung Lee (ljy@mobis.co.kr)

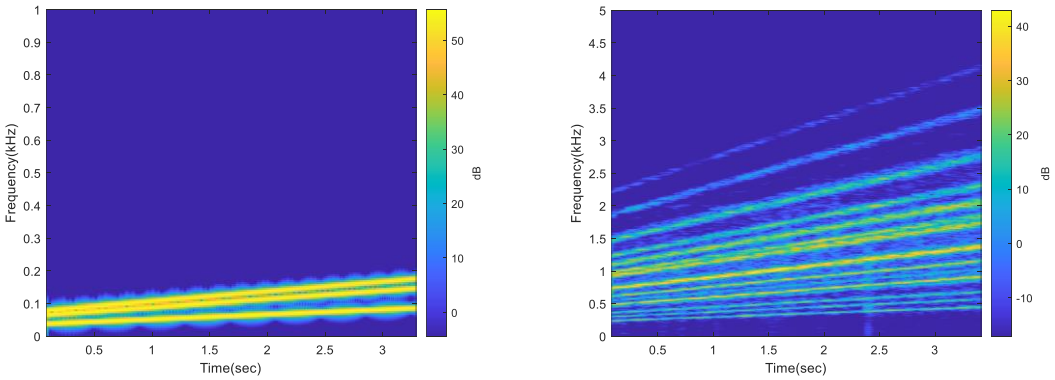
ABSTRACT

The virtual engine sound can generate various sounds from an internal combustion engine to a spacecraft sound using sources. However, since it is hard to analyze sources, synthesizers, and control variables, it is difficult to produce a target sound. This paper proposes a virtual engine sound generation method based on a transformer to design sounds suitable for the target vehicle. The proposed method learns a virtual engine sound generator using data acquired according to the driving situation. Moreover, the transformer is used to alleviate repeatability. Therefore, this method can reproduce virtual engine sounds with complex structures and change styles.

1 Introduction

Recently, as engine efficiency has improved, the amount of energy consumed to produce sound has decreased. It leads to a decline in the emotional quality of engine sound. The ESE (Engine Sound Enhancement) system provides high-quality engine sound to drivers by outputting similar sounds from speakers [1]. Since the engine generates power by repeating processes such as intake, compression, explosion, and exhaust, the frequency is determined according to the number of engine rotations. Therefore, a tone generator can reinforce the engine sound to synthesize a sine wave with fundamental frequency and harmonic components depending on the engine speed [2]. The engine sound gives the driving state information to the driver and improves driving immersion, so virtual engine sound is also being introduced in electric vehicles. Since electric vehicles have quiet motor driving sounds, the virtual engine sounds can be extended from traditional internal combustion engines. A wavetable synthesizer is introduced, which generates sound by repeatedly playing sound sources with arbitrary waveforms. In addition, a unique virtual engine sound can be generated by using a granular synthesizer, which uses sound sources by dividing them into short sections [3][4].

A synthesis method using sound sources with a high degree of freedom can express sounds in various contexts. By changing the sound source, synthesis method, and control variables, it is possible to generate from internal combustion engine sounds to spaceship sounds. However, there is no means to describe engine sounds like sheet music. Thus, a sound suitable for the target vehicle is designed to change the style of the other vehicles. Since the ESE system synthesizes sinusoidal waves with a single frequency, the number of tones, orders, and gains can be analyzed easily by using the spectrogram according to the driving situation as shown in Fig. 1. However, it is difficult to reproduce electric vehicle virtual engine sounds by identifying the synthesis method, sound source, and control variables from the synthesis results. A sampling method has been proposed to recreate engine sounds with complex configurations. This method creates sound by recording engine sounds from other vehicles and then playing back sounds appropriate for the driving situation. However, since recording is done by



(a) Tone generator (b) Wavetable synthesizer

Figure 1. Virtual engine sound spectrogram during acceleration

rotation per minute (RPM), it is easy to feel disconnected. A large amount of storage space is required to store data for each driving situation. Also, it is hard to change style. With the development of deep learning technology, data-driven high-quality audio generation methods have been proposed. Generative adversarial network (GAN)-based audio generation methods use spectrograms of audio signals as images or one-dimensional convolution to construct a generator and discriminator. Although this method can make new engine sounds using virtual engine sound data, it cannot generate sounds according to driving situations. The variational auto encoder (VAE) based audio generation method analyzes the probability distribution of data acquired by encoders. Thus, it generates data with a distribution like the input data. Although this method can generate similar high-quality audio, it has the problem of requiring input data to produce new sounds tailored to the driving situation. Moreover, methods for reproducing wavetables or analyzing grains and control variables have been proposed. However, similar sounds cannot be generated when only the driving situation is given.

In this paper, we propose a transformer-based virtual engine sound generation method to design a sound suitable for the target vehicle. The proposed method produces a virtual engine sound, such as the acquired data according to vehicle speed, torque, RPM, and accelerator pedal sensor (APS). Drivers feel bored or dizzy when the virtual engine sound is repeated. A transformer is used to alleviate the repeatability of the generated virtual engine sound. Thus, even if the same driving situation is repeated, the sound is deformed. In general, the target virtual engine sound is similar to the existing one but has different emotions and styles. Audio style refers to the type or texture of sound, but it is difficult to clearly define it. Therefore, when style transfer based on spectrogram is used in the image area, the result is a mixing of two sounds rather than a change in characteristics. The proposed method constructed a style evaluation network that classifies genres or emotions and used the output features of this network to determine the similarity of audio styles. Therefore, style is applied to the reconstructed virtual engine sound by learning to make the characteristics of the style sound source and the characteristics of the conditional transformer output similar.

This paper is structured as follows. The proposed method is described in section 2, and section 3 presents the experimental results. Finally, we conclude in section 4.

2 Methods

2.1 Related Works

Sampling Method: To generate engine sounds in electric vehicles, a method of recording and playing back the engine sounds of internal combustion engine vehicles has been proposed. This method records from 1500 RPM to 9000 RPM while keeping the engine RPM constant, and then plays it back according to the APS value. This method reproduces engine sounds recorded for each RPM section, so the sound does not naturally continue during acceleration and a sense of disconnection is felt. In addition, a lot of memory is used to store engine sounds, and it is difficult to create unique sounds.

WaveGAN: GAN is a generation model consisting of a discriminator and a generator. It receives random numbers, creates an image in the generator, and determines whether it is a synthetic image or not in the discriminator. WaveGAN is an application of GAN to audio. It uses a random shuffle to alleviate the checkboard phenomenon that occurs during

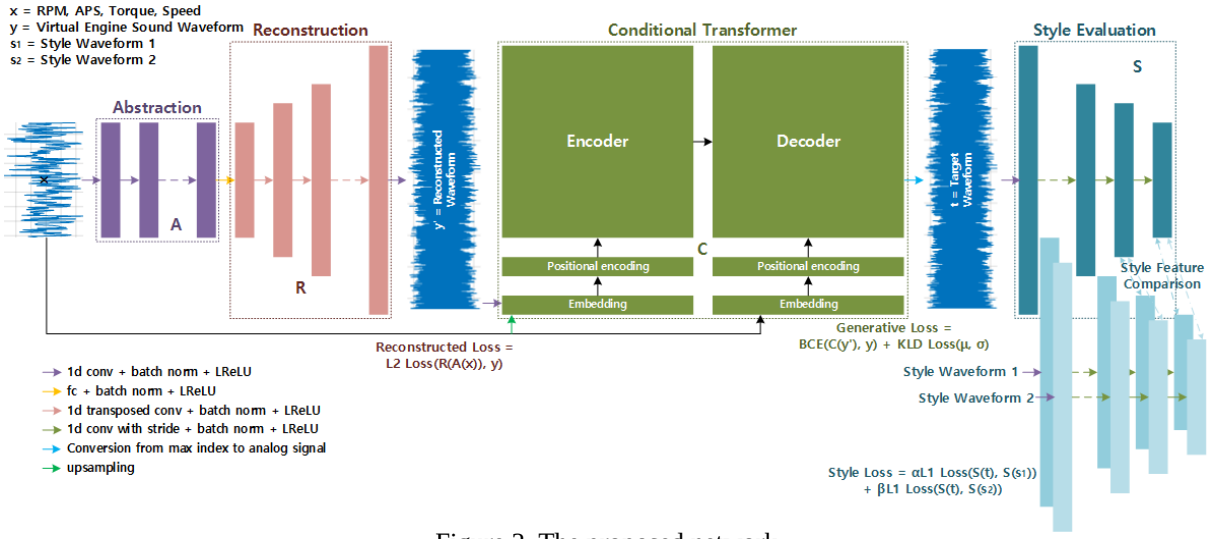


Figure 2. The proposed network

upsampling. Thus, it can generate at a faster rate than autoregression methods such as WaveNet [5]. However, it is difficult to make a natural virtual engine sound suitable for the driving situation.

RAVE: The VAE encodes the input data into a latent space with a normal distribution and then generates new data by obtaining a random value. However, since the human ear is sensitive to both microscopic and macroscopic changes in audio waveforms, audio generated by VAE has a low emotional quality. Realtime audio variational autoencoder (RAVE) can generate high-quality audio by controlling reconstruction fidelity and expression compression through secondary learning [6]. However, since sound cannot be generated using only driving status signals, it is not suitable for generating virtual engine sounds.

Style Transfer: In the image, style defines the characteristics of the entire image, not just a small area. Therefore, features are extracted using the learned classification network, and learning is performed so that the correlation between regional traits is similar. To transfer styles in audio, conversion is performed using the spectrogram of the signal [7]. In audio, style represents differentiated quality, style, and form, and is classified by era, composer, performer, genre, and texture. However, because the spectrogram has strong expressive power for the entire feature rather than the local waveform change of the signal, it is difficult to match the gram matrix result of the feature map with the style. Therefore, the style conversion result is like the synthesis of two signals rather than a change in shape or form.

2.2 Proposed Method

Problem Setting: The virtual engine sound of the electric vehicle can express various sounds depending on the sound source, synthesis method, and control variable. However, since virtual engine sound does not have the means to describe sound like sheet music, it is difficult to efficiently generate engine sound that meets the intention. Even if the planner describes the target virtual engine sound in detail, the meaning understood by the composer may be different. Therefore, the sound created by selecting the synthesis method and composing the sound source may be different from the planner's intention. To reduce trial and error, a method of changing the style based on the engine sound of a specific vehicle is employed. However, it is difficult to identify synthesizers or sound sources from the synthesis results. It is even difficult to confirm whether it is possible to imitate the virtual engine sound of a specific vehicle by adjusting the sound source and control variables based on a specific synthesis method.

Abstraction & Reconstruction: The sampling frequency of the driving state signal x is 10 Hz to 50 Hz, but the sound source y is 48 kHz. Therefore, it is necessary to increase the size of the feature map by using upsampling to match the driving status signal. Abstraction, A, increased the level of abstraction by increasing the number of layers and channels with the same size as the driving state signal to increase computational efficiency. In addition, the size of the feature map is increased to have the same length as the waveform using one-dimensional transposed convolution in the reconstruction module, R. In the

intermediate stage, the leaky rectified linear unit (LReLU) is used as an active function. At the end, a hyperbolic tangent is employed. In addition, to receive the driving state signal and generate the same signal as the virtual engine sound of other vehicles, it was learned using a loss function as shown in Equation 1.

$$reconstruction\ loss = \sum \left(y - R\left(A(x)\right)\right)^2 \quad (1)$$

Conditional Transformer: When the driving status is the same, the output of R has the same value. Therefore, because repetitive waveforms are output, the driver may feel dizzy or bored. To increase the naturalness of the virtual engine sound, the output should change slightly while maintaining the texture in the same driving situation. The proposed method uses a conditional transformer, C, to provide natural waveform changes while maintaining the theme of the generated waveform like a variation. The conditional transformer maintains the waveform according to the driving state signal, but the output changes little by little depending on the conditions generated by random numbers. Therefore, it is possible to output natural variations, unlike amplitude randomization, where the amount of change is determined using probability values. Moreover, to force small changes, content loss is used as given in Equation 2 to preserve the original sound when converting audio styles.

$$content\ loss = -\sum \begin{matrix} y' \log C(y', x) + \\ (1 - y') \log\left(1 - C(y', x)\right) \end{matrix} \quad (2)$$

Here, y' is $R\left(A(x)\right)$.

Style Evaluation: Style cannot be mathematically defined because it represents an abstract type or form. In the image field, it is expressed using the correlation of local features of the classification network and evaluated qualitatively based on visual similarity. If a spectrogram is used in audio, the style can be expressed in the same. However, since there is no local feature transformation, the results are like simple synthesis. The proposed method uses a music genre classification network, S, to express audio styles based on features of the network. Therefore, a loss function is defined as in Equation 3 to have similar features between the output of the conditional transformer and the style sound source.

$$style\ loss = \begin{matrix} \alpha \sum |S(t) - S(s_1)| \\ +\beta \sum |S(t) - S(s_2)| \end{matrix} \quad (3)$$

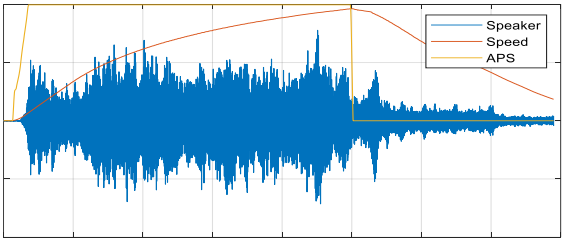


Figure 3. Signals included in dataset

Here t is $C(y')$. To enable continuous style changes, constant values, α and β, are used.

System Behavior: The proposed method builds a database by acquiring driving state signals and sounds from existing vehicles. It learns the network to design virtual engine sounds suitable for the target vehicle. It uses the abstraction and reconstruction modules to restore the virtual engine sound of the existing vehicle from driving state signals such as vehicle speed, torque, RPM, and APS, as shown in Fig. 2. The abstraction module abstracts the input driving state signal and extracts features. In addition, it is learned to generate the same output as the existing vehicle signal by performing upsampling through the reconstruction module. When there is no change in driving conditions, the same virtual engine sound is repeatedly generated, which reduces emotional quality. The proposed method used a conditional transformer to make variation sounds and change the style. Style is defined using a pre-trained style evaluation network. Then, the conditional transformer is learned to generate output sounds with similar features to the style sound source. Therefore, using the proposed method, it is possible to create a synthesizer that generates target virtual engine sounds described by existing engine sounds and style sound sources.

3 Experiment Results

Dataset: The virtual engine sound generates sounds according to driving signals and outputs them through speakers. To perform supervised learning, data was acquired for 10.7 hours from Hyundai KONA vehicle under various driving conditions. A sampling frequency of 44.1 kHz is used, and the acquired data are vehicle speed, APS, and speaker output as shown in Fig. 3.

Quantitative Analysis: The sound waveform is quantized to 9 bits, and the next sample is inferred using 2048 samples. The accuracy is shown in Table 1.

Top-1 accuracy	Top-5 accuracy
73.3%	99.9%

Table 1. Waveform inference accuracy

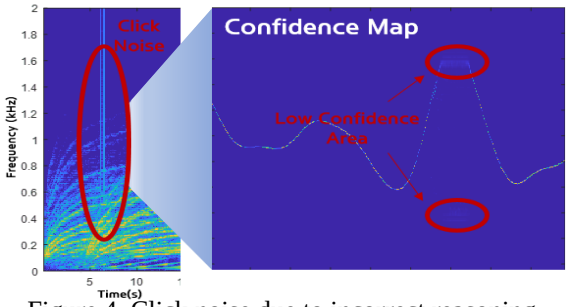


Figure 4. Click noise due to incorrect reasoning

Errors of 2 bits or less that occur intermittently in analog waveforms are difficult to perceive by the human ear. However, as shown in Figure 4, fluctuations may occur in low-confidence sections, which may cause click noise. The proposed method solves this problem by limiting the amount of change in the new inference value based on the previous one.

Qualitative Analysis: To verify the proposed method, the KONA sounds are learned and reproduced during acceleration/deceleration driving situations. The virtual engine sound of the Kia EV9 vehicle is used as a style sound source. Moreover, a change amount constraint is imposed. Figure 5 shows that it generates a virtual engine sound similar to KONA but with a changed style.

4 Conclusions

In this paper, a transformer-based virtual engine sound generation method is proposed to design a sound suitable for the target vehicle by using the existing engine data. Engine sounds are developed by selecting a target sound, but it is not easy to identify the synthesis method, sound source, and control variables. However, the proposed method generates similar engine sounds by learning using data acquired in various driving situations. Repeatability is alleviated using a conditional transformer by changing the sound slightly. In addition, defining the style using the evaluation network, can change the style of the generated sound. To verify the proposed method, a database is constructed by acquiring driving data and sounds from an actual vehicle. As a result of learning, it is confirmed that the generated sound similar to the actual one. Therefore, the proposed method can also simulate a virtual engine sound generator of a target vehicle with a complex internal structure.

References

[1] R. Schirmacher, "Active Design of Automotive Engine Sound," Journal of the Audio Engineering Society, 2002.

[2] M. Dongki, P. Buhm, P. Junhong, “Artificial Engine Sound Synthesis Method for Modification of the Acoustic Characteristics of Electric Vehicles,” Shock and Vibration, 2018.

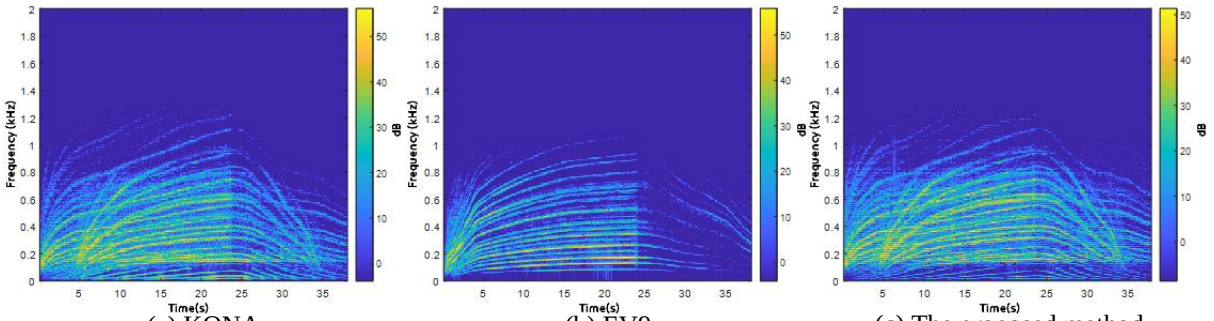
[3] R. Bristow-Johnson, " Wavetable Synthesis 101, A Fundamental Perspective," Journal of the Audio Engineering Society, 1996.

[4] J. Jagla, J. Maillard and N. Martin, "Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-And-Add method," The Journal of the Acoustical Society of America, 2012.

[5] Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial Audio Synthesis. ArXiv. 1802.04208.

[6] Caillon, A., & Esling, P. (2021). RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. ArXiv. 2111.05011.

[7] Grinstein, E., Duong, N., Ozerov, A., & Pérez, P. (2017). Audio style transfer. ArXiv. https://doi.org/10.1109/ICASSP.2018.8461711.



(a) KONA (b) EV9 (c) The proposed method

Figure 5. Virtual engine sound generation result (content: KONA, style: EV9)