Proceedings of the Canadian Society for Mechanical Engineering International Congress
32nd Annual Conference of the Computational Fluid Dynamics Society of Canada
Canadian Society of Rheology Symposium
CSME-CFDSC-CSR 2025
May 25–28, 2025, Montréal, Québec, Canada

GENERATING REALISTIC ENGINE SOUNDS FOR SIMULATED CONSTRUCTION EQUIPMENT

Travis Wiens
Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, Canada
t.wiens@usask.ca

Abstract—Computer simulations of construction equipment is important for training and research applications. For example, one recent research topic is to understand how human operators of equipment react to the sounds of the machine, which can give feedback on forces, velocities, power levels, pressures, etc. This type of research is often performed in simulation for cost and safety reasons, but requires a method of generating accurate sounds. This paper presents a dataset of audio recordings from a backhoe loader at varying engine speeds and hydraulic system loadings as well as a preliminary attempt at generating synthetic audio that sounds plausible to human perception.

Keywords- audio, engine, hydraulic, fluid power, audio synthesis

I. INTRODUCTION

The combination of a diesel engine powering a hydraulic system is common across mobile equipment including agricultural, mining and construction equipment. These systems exhibit very high power and force densities, but existing systems can have low energy efficiencies. Love [1] estimated that the efficiency of hydraulic systems across industry in the US have a hydraulic system energy efficiency of 21%. This efficiency does not include the low thermodynamic efficiency of diesel engines, which can range from 43-49% [2]. Combined, a large amount of energy is wasted, with a corresponding unnecessary emission of greenhouse gases.

There is considerable active research on improving the energy efficiency of mobile equipment through hardware system design, but the human operator is also a considerable contributor to inefficient operation. For example Frank et al. [3] showed that a poor operator of a wheel loader can use 200-300% more fuel than a good one, and even good operators vary considerably over the day.

One research aspect to be studied is the influence of audio feedback on the performance and energy efficiency of human operators. By listening to the sounds produced by the engine and hydraulic system, the operator can gain information about the system pressures, forces and power levels that are not otherwise available. It has been hypothesized that this can be an important factor related to efficient operation of mobile equipment. The author's laboratory has constructed simulators intended to test this hypothesis [4] but it is challenging to generate accurate and plausible engine sounds that convey the load levels of the engine to the operator while not distracting the operator.

A number of conventional sound synthesis methods were attempted with disappointing results. For example wave table synthesis [5] was attempted but this required a large number of repetitive wave forms at each operating point in order to sound natural and was only applicable at discrete operating points with limited ability to interpolate between those points. We attempted to use longer samples of prerecorded data to ensure that the sound had realistic variation from combustion cycle to combustion cycle but this also suffered from issues with seamlessly moving from one operating point to another.

This paper presents an attempt to solve this problem by first generating a large library of recordings of a real engine at varying speed and load levels, segmenting the recording into cycles of a single camshaft rotation, and then attempting to recreate the sounds by generating waveforms that are statistically similar to the library recordings. This paper will also present some preliminary results from human listeners quantifying their ability to determine the difference between real and synthetic sounds.

II. METHODOLOGY: AUDIO RECORDINGS

A dataset of audio recordings was recorded from a John Deere 410D Backhoe Loader, as shown in Figure 1. This machine is powered by a four cylinder diesel engine with a seven-bladed fan connected to the crankshaft via a belt drive with nominally equal pully diameters, and a nine piston axial piston hydraulic pump also directly connected to the crankshaft. An external relief valve was connected to the loader hydraulic system's auxiliary output to provide an adjustable load for the engine. The hydraulic system was limited to 3000 psi (20.7 MPa), less than the machine's rated 3625 psi (25 MPa) so the engine could not be fully loaded.

A Zoom H4n Pro portable digital audio recorded was used to record the sound of the machine. The recorder with its dual

internal microphones were placed on a stand outside of the cab, near the machine. An external microphone (Shure SV100) was located in the cab of this machine, near the operator's ear position (no operator was inside the cab at the time of recordings to reduce variability in the acoustic environment). Although the exterior recordings are available, this paper will use the cab interior data as this is most representative of the operator's experience. Recordings were taken as .wav files with 16 bits per sample at a rate of 44.1 kHz.

Recordings were taken for a range of engine speeds and load pressure levels. For each speed, the pressure valve was set to 0 psi to unload the engine and the engine speed control dial was used to set the crankshaft speed as close as possible to the desired setpoint, $\widehat{N}$, using the machine's tachometer dial. The pressure was then increased to the desired load pressure, P, (monitored via an analog pressure gage) without adjusting the engine speed dial, allowing the engine speed, N, to reduce under load. At each operating point, the system was allowed to stabilize and then 30-120 s of audio was recorded. Further details on the recording dataset are available from [6].

Figure 1: John Deere 410D backhoe loader used in this study.

III. Analysis: Audio Recordings

A. Preprocessing

For each recording, we first listened to the recording to ensure that there were no anomalies within. Then we determined the average engine speed for the recording. We calculated the circular autocorrelation [7] for the recording and looked for a peak within the range of the setpoint engine speed +/- 20%. The dataset was then divided into segments, each corresponding to one rotation of the engine's camshaft (the camshaft rotates at half the crankshaft speed). This will allow each cylinder of the engine to complete one combustion cycle and the sound attributable to the combustion events would be ideally aligned and repeatable for each segment. The data was first segmented using the average engine speed and then fine adjustments were made to account for the slightly varying engine speed from cycle to cycle. This was achieved by identifying the peak in the correlation between the segment and the average of all segments at that operating point, identifying the relative time shift for that segment. The adjusted segment length was then shifted in time to align with the average segment and resampled using piecewise cubic spline interpolation so that each segment in the recording has the same number of samples.

B. Audio Analysis

This section includes an analysis of the basic characteristics of the recorded audio. Although we have recordings from a wide variety of operating points, will concentrate on recordings from three: 1) $\widehat{N}$ =930 rev/min setpoint engine speed and P =0 psi hydraulic load, representing an unloaded idle condition, 2) 2200 rev/min and 800 psi representing full speed and light load, and 3) 2200 rev/min and 3000 psi representing full speed and high load. Figure 2-4 shows the distribution of time series data, x, for the three operating points, showing 20 randomly selected time series in grey and the mean +/- 1 standard deviation bounds in black. Each shows one camshaft revolution including a single combustion event from each of the four cylinders. At idle (Figure 2) the four individual combustion events are clearly visible, while this is not the case for the full speed recordings, which have comparatively more acoustical content contributed by the cooling fan and hydraulic pump. Note the differing magnitudes of the recordings, with the fully loaded case being the loudest (as expected).

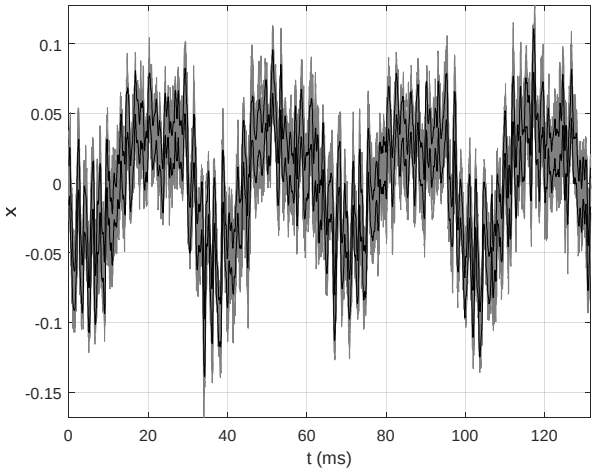

Figure 2: Time series recording for a single camshaft revolution at $\widehat{N}$=930 rev/min setpoint engine speed and P=0.

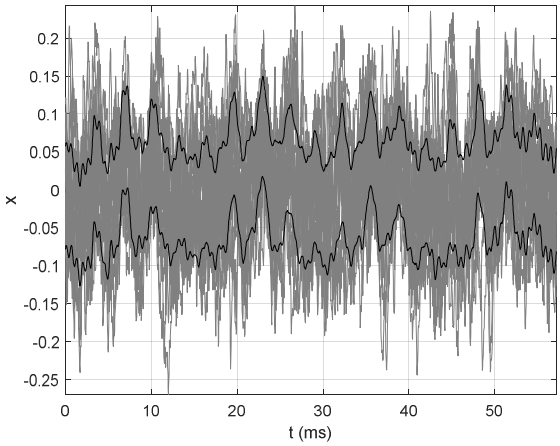

Figure 3: Time series recording for a single camshaft revolution at $\widehat{N}$=2200 rev/min setpoint engine speed and P=800 psi.

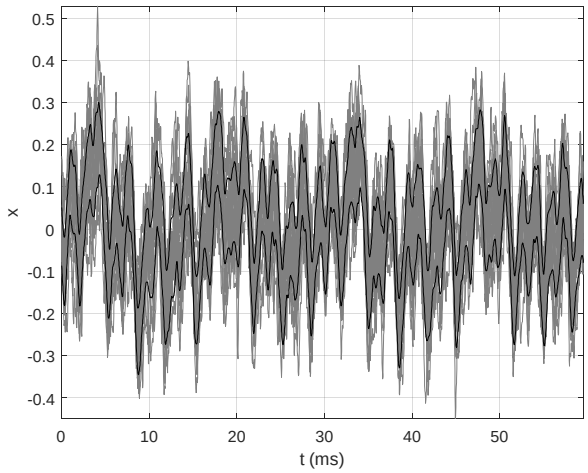

Figure 4: Time series recording for a single camshaft revolution at $\widehat{N}$=2200 rev/min setpoint engine speed and P=3000 psi.

Figures 5-7 show the A-weighted frequency response, with the frequency normalized by the cam shaft frequency, f_c. Again, this shows 20 randomly selected cycles in grey and the mean +/- 1 standard deviation in black. Also shown are the first five harmonics of some identifiable sources of sound: the combustion event that occurs four times per cam cycle, the nine piston hydraulic pump (18 events per cam rotation) and the seven blade engine cooling fan (14 events per cam rotation). Each of these sources show significant peaks in the frequency response but there is significant sound power in between these peaks as well.

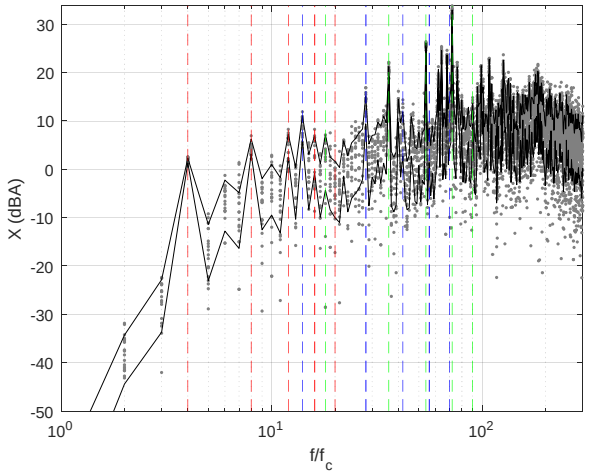

Figure 5: Frequency response for $\widehat{N}$=930 rev/min setpoint engine speed and P=0. The first five harmonics of the combustion cycle, hydraulic pump pistons and fan blades are shown in red, green and blue.

The relative phase of the ten strongest frequency components is shown in Figure 8-10, with the normalized frequency, f/f_c, labelled for the strongest components. This plot is the real and imaginary part of the Fourier transformed data, with the distance from the center indicating the magnitude and the angle counterclockwise from the x axis indicating the phase. Each color indicates a different frequency component, with all cycles

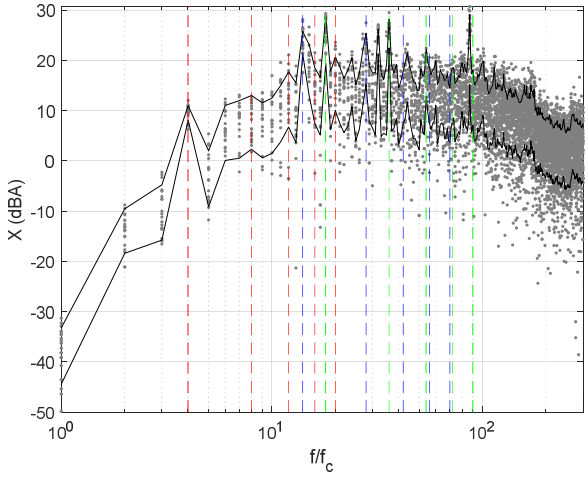

Figure 6: Frequency response for $\widehat{N}$=2200 rev/min setpoint engine speed and P=800 psi. The first five harmonics of the combustion cycle, hydraulic pump pistons and fan blades are shown in red, green and blue.

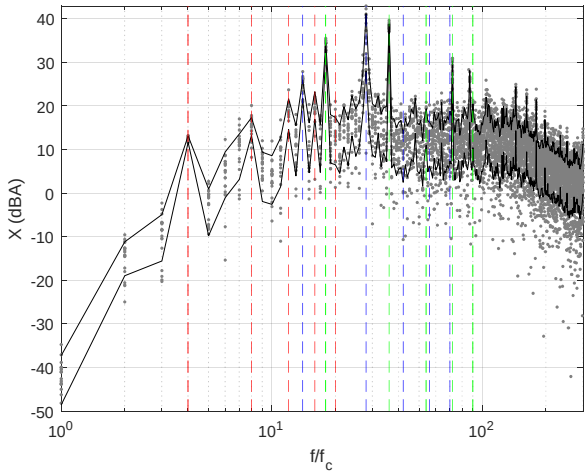

Figure 7: Frequency response for $\widehat{N}$=2200 rev/min setpoint engine speed and P=3000 psi. The first five harmonics of the combustion cycle, hydraulic pump pistons and fan blades are shown in red, green and blue.

plotted here. These figures demonstrate the variability in both magnitude and phase. For example, in Figure 8, the cluster for f/f_c=4 is relatively tight in both magnitude and phase, while the cluster for $f/f_c = 72$ is consistent in magnitude but has a wider phase variation. In Figure 9 most of the clusters have considerable variability while when the engine is loaded (Figure 10) the clusters become more well defined. This may be due to the oscillatory nature of the engine governor when not fully loaded. Also note that in Figure 10, the highly loaded engine exhibits tight clusters for $f/f_c = 4$ (engine combustion) and 18 (hydraulic pump), while the fan frequency (f/f_c=14 and a harmonic at 28) have high magnitude but very large variation in phase. There is also considerable content at f/f_c=27 in this case; the variable phase and bleed into lower frequency bands may indicate that the belt connecting the fan to the engine is slipping slightly or that the pullies are slightly different diameters.

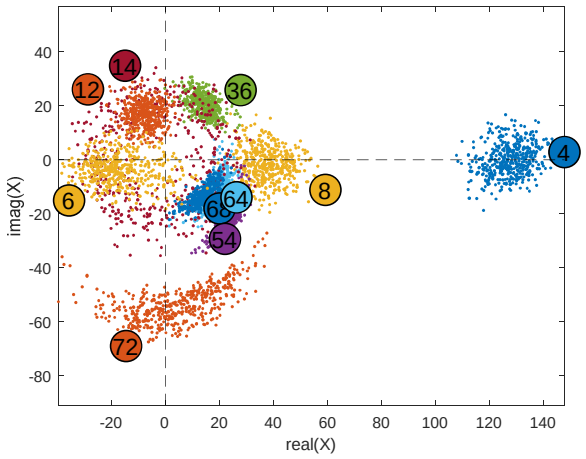

Figure 8: Phase portrait of the ten strongest components in the frequency response for $\bar{N}$=930 rev/min setpoint engine speed and P=0 psi.. Numbers indicate the frequency relative to the camshaft frequency.

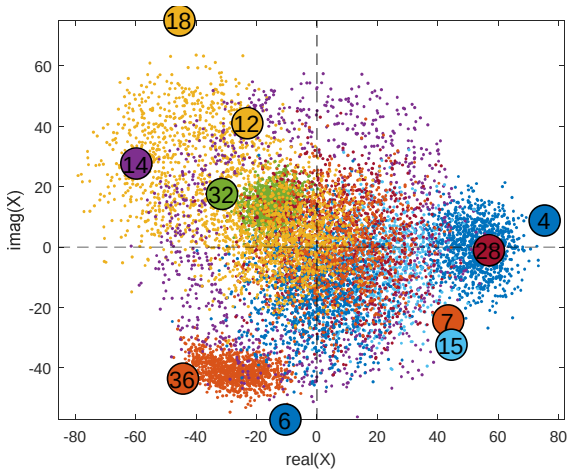

Figure 9: Phase portrait of the ten strongest components in the frequency response for $\bar{N}$=2200 rev/min setpoint engine speed and P=1000 psi.. Numbers indicate the frequency relative to the camshaft frequency.

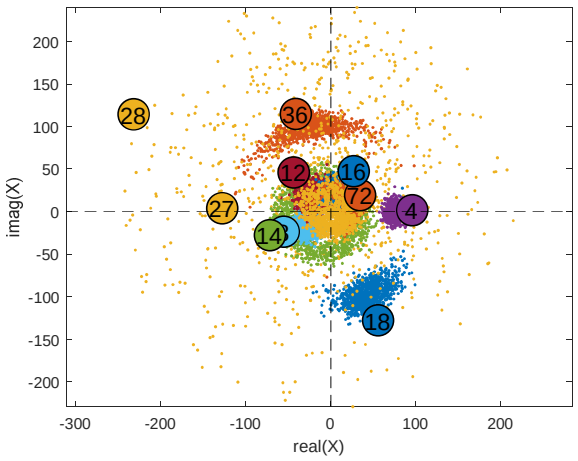

Figure 10: Phase portrait of the ten strongest components in the frequency response for $\bar{N}$=2200 rev/min setpoint engine speed and P=3000 psi.. Numbers indicate the frequency relative to the camshaft frequency.

III. Methodology: Audio Synthesis

In this section we present several methods of synthesizing sounds. In each case we present perceptual listening tests to determine whether a human can determine the difference between synthesized and real audio. This was implemented via a forced choice triangle test or AAB test [8], where three short samples of sound (32 cam rotation cycles) were presented to the listener. Two would be real audio and one synthesized (or one real and two synthesized) and the listener is asked to identify the different sample. In each case we presented the data in five groups of three repetitions of the same method and operating point, with different sections of the real sound and newly generated synthetic data. Listeners were presented with a interface which would allow them to listen to each sample as often as required to make a choice. Listeners were required to wear headphones to ensure accurate sound representation.

Each is test statistically analyzed using the single-tailed binomial test under the null hypothesis that the samples are the same and any correct responses can be attributed to chance with a probability p.

We tested five methods of audio synthesis to evaluate the synthesis methods, but also to understand the importance of various features of the sound when perceived by human listeners.

A. *Single Cycle Repetition*

This is the simplest method of audio synthesis trialed. A single cycle of the data library is randomly selected and is repeated 32 times to generate the sound sample. This repetitive sound can be used to evaluate how important cycle-to-cycle variation is to listeners and whether a simple method like wave table synthesis can be used to represent the data

B. *Shuffled Cycles*

In this case we selected 32 consecutive cycles from the data library and reassembled them into a sound sample in random order (with no repetition). This indicates whether the order of the samples is important to human perception or whether there are features of a cycle waveform that must have specific qualities related to the previous cycle.

C. *Reduced Shuffled Cycles*

This is similar to the above method but a reduced number of cycles, N_c, are selected and then reassembled in random order, requiring repetition. For $N_c = 1$ this converges to Method A and $N_c = 32$ converges to Method B. This can be used to indicate how much diversity in cycle generation is required before the sound becomes repetitive.

D. *Statistical Time Domain*

In this test, the mean and standard deviation of the time signal (as shown in Fig 2-4) were used to generate new data points at each point in the time domain. The results for this test would indicate whether the data can be adequately represented by an independent identically distributed (IID) normal distribution.

E. Statistical Frequency Domain

In this test, the mean and standard deviation of the signal in the frequency domain (Fig 5-7) is used to generate new frequency domain samples, which are then converted into the time domain using the inverse Fourier transform.

F. Generation Using a Variational Autoencoder

This test was similar to the previous statistical frequency domain sound generation, but rather than treating each frequency magnitude and phase as normally and independently distributed quantities, a variational autoencoder (VAE) [9] was used to attempt to model and recreate the distribution including non-normal effects and correlations between quantities. Due to time constraints, we were only able to perform this test for Operating Points 2 and 3.

The VAE was set up with ReLU [10] fully connected layers with 100 and 400 units in the encoder which then narrowed to 16 in the latent encoding layer. The decoder was similarly sized. Matlab's implementation of the ADAM optimizer [11] was used to train the network, with a learning rate of 0.0032 and trained for 100 000 epochs. This was run on a consumer desktop computer with an AMD Ryzen 3950X CPU, 32 GB RAM and a GeForce RTX 2080Ti GPU.

IV. Results: Audio Synthesis

A. Single Cycle Repetition

In this test we compared samples of a single repeated cycle to a recording containing multiple sequential cycles. The results are shown in Table 1. Clearly the human listeners had no trouble differentiating between the real recording and the repeated single cycle, which was very repetitive. This indicates that some measure of variation from cycle to cycle must be included in any realistic sounding synthesis method.

Table 1: Perceptual Testing Results for a single cycle repetition. A total of N_r=10-15 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	0.9	0.00
2	2200	800	1.0	0.00
3	2200	3000	1.0	0.00

B. Shuffled Cycles

We then tested the difference between a normal recording and one where the individual cycles were randomly reordered. The results are shown in Table 2. In this case, the listeners were not able to reliably determine the difference for operating Points 1 and 2 (i.e. p>0.05). Although the panel size (N=2) was small, this indicates that the history of the cycles is not critical at these operating points and each cycle can be generated independently, without requiring information on the previous cycle. There was, however, an ability to detect the difference for Operating Point 3. After discussing with the panel, a warble at a specific frequency was identified which corresponded to the fan's frequency of f/f_c =14 and the first harmonic at f/f_c=28. As noted earlier, the phase angle at these two frequencies are distributed throughout the angular range, which is believed to be due to the fan drive slipping leading to a slightly lower frequency, which would no longer align with the cam shaft frequency. We tested this by removing these two frequencies (in the Fourier domain) and repeating the perceptual test, with results shown in Table 3. Now the participants could not reliably detect the difference between the real and the shuffled sounds. This indicates that the fan-related frequencies may need to be generated using a separate method for most accurate simulation.

Table 2: Perceptual Testing Results for shuffled cycles. A total of N_r=10 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	0.6	0.07
2	2200	800	0.6	0.07
3	2200	3000	0.9	0.00

Table 3: Perceptual Testing Results for shuffled cycles with fan frequencies removed. A total of N_r=15 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	0.47	0.20
2	2200	800	0.40	0.38
3	2200	3000	0.20	0.20

C. Reduced Shuffled Cycles

We repeated the above test, but instead of using each cycle only once (32 total cycles), we selected a reduced number of cycles with repetition (again randomly selected). As shown in Table 4, the listeners were not able to reliably discriminate between real and synthetic sounds for the case of Operating Point 1 as long as N_c=16 samples were used. The other two cases required the full 32 samples from Method B.

Table 4: Perceptual Testing Results for shuffled cycles, selected from N_c cycles. A total of, N_r=15-20 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	N_c	Fraction Correct	p
1	930	0	4	0.87	0.00
2	2200	800	4	0.75	0.00
3	2200	3000	4	1.00	0.00
1	930	0	8	1.00	0.00
2	2200	800	8	1.00	0.00
3	2200	3000	8	1.00	0.00
1	930	0	16	0.60	0.31
2	2200	800	16	0.90	0.00
3	2200	3000	16	1.00	0.00

D. Statistical Time Domain

In this test, the mean and standard deviation of the time signal were used to generate new normally distributed samples. As shown in Table 5, the listeners were able to reliably differentiate the real and synthetic sounds; in all cases p<0.05. This is believed to be due to the fact that the random generation includes more high frequency content, perceived as white noise overlaid on the signal.

Table 5: Perceptual Testing Results for synthesized sounds using time-domain statistics. A total of N_r=15 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	0.93	0.00
2	2200	800	0.67	0.01
3	2200	3000	0.87	0.00

E. Statistical Frequency Domain

In this test the mean and standard deviation of the frequency signal phase and magnitude were used to generate new samples in the frequency domain. As shown in Table 6, the panelists were able to reliably differentiate the synthetic sounds, likely due to an inaccurate generation of the phases of the various frequency components and their correlation with each other.

Table 6: Perceptual Testing Results for synthesized sounds using frequency-domain statistics. A total of N_r=15 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	0.93	0.00
2	2200	800	1.00	0.00
3	2200	3000	1.00	0.00

F. Generation Using a Variational Autoencoder

In this test a variational autoencoder (VAE) was used to attempt to model and recreate the statistical distribution including non-normal effects and correlations between quantities. As shown in Table 7, the listeners were not able to reliably differentiate between the real and synthetic sounds, indicating that this method shows promise as a method of generating plausible engine sounds.

Table 7: Perceptual Testing Results for synthesized sounds using frequency-domain statistics. A total of, N_r=10 choices were requested at each operating point.

Operating Point	$\widehat{N}$ (rev/min)	P (psi)	Fraction Correct	p
1	930	0	n/a	n/a
2	2200	800	0.40	0.44
3	2200	3000	0.50	0.21

V. CONCLUSIONS

The results presented here give initial indications that:

1) The variation in sound from cycle to cycle is significant and individual samples of sound are not sufficient to play repeatedly

2) If a discrete number of samples played in random order would be used to generate the sound, the number of required samples is quite large: probably at least 16 per operating point

3) Generating samples statistically assuming a normal IID distribution in either the time domain or frequency domain is not convincing

4) A variational autoencoder appears to show promise as a method of generating sounds that considers the non-normal and not IID nature of the data.

These conclusion should be taken as very preliminary. The number of listeners was small and may not be representative of the larger population. A study is ongoing that will include a larger panel of listeners.

While the variational autoencoder appears to be successful at generating sounds at a single operating point, this does not satisfy the objectives of this project. We wish to be able to generate sounds at any operating point without requiring a very fine grid of recordings to use as training data for a large number of separate VAEs. Ongoing future work will study the potential for Conditional Variational Autoencoders (CVAEs) [12], such that a single CVAE network can take inputs of arbitrary engine speed and load and generate convincing sounds.

ACKNOWLEDGMENT

The author would like to acknowledge Meagan Lee and Doug Bitner for their work recording the original dataset. This work was financially supported by the NSERC USRA program.

This study was approved by the University of Saskatchewan Behavioral Research Ethics Board, certificate 4998.

REFERENCES

[1] L. J. Love, E. Lanke, and P. Alles, "Estimating the impact (energy, emissions and economics) of the US fluid power industry," *Oak Ridge National Laboratory, Oak Ridge, TN*, 2012, Accessed: Oct. 02, 2024. [Online]. Available: https://info.ornl.gov/sites/publications/files/pub28014.pdf

[2] R. Giannelli, E. Nam, and A. Arbor, "Medium and heavy duty diesel vehicle modeling using a fuel consumption methodology," *National Vehicle Fuel and Emission Laboratory, US Environmental Protection Agency: Ann Arbor, MI, USA*, 2004.

[3] B. Frank, L. Skogh, and M. Alaküla, "On wheel loader fuel efficiency difference due to operator behaviour distribution," in *2nd International Commercial Vehicle Technology Symposium, CVT*, 2012, pp. 1–18.

[4] T. Wiens, M. Klarkowski, and N. Zahabi, "Development of a Physical Analog Excavator for Studies in Interactions Between Hydraulic Equipment and Human Operators," in *BATH/ASME 2020 Symposium on Fluid Power and Motion Control*, American Society of Mechanical Engineers, Sep. 2020. doi: 10.1115/fpmc2020-2771.

[5] C. Roads, *Computer Music Tutorial*, vol. 170. MIT Press, 1996.

[6] M. Lee, "Undergraduate Research Project Report: Generating Plausible Engine Sounds." Department of Mechanical Engineering, University of Saskatchewan, Sep. 02, 2021.

[7] T. Wiens, *Fast Circular (Periodic) Cross Correlation*. (2009). Accessed: Oct. 01, 2024. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/24342-fast-circular-periodic-cross-correlation

[8] M. O'Mahony, *Sensory Evaluation of Food: Statistical Methods and Procedures*. New York: Routledge, 2017. doi: 10.1201/9780203739884.

[9] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[10] K. Fukushima, "Visual feature extraction by a multilayered network of analog threshold elements," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 4, pp. 322–333, 1969.

[11] P. K. Diederik, "Adam: A method for stochastic optimization," *arXiv preprint*, 2014.

[12] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.