GAN Based Audio Noise Suppression for Victim Detection at Disaster Sites with UAV

Chinthaka Premachandra, *Senior Member, IEEE,* and Yugo Kunisada

***Abstract*— This paper presents a noise suppression system designed for unmanned aerial vehicles (UAVs). Searching for people using robots is expected to become a useful tool for saving lives during disasters. In particular, because UAVs can collect information from the air, there has been much research in rescue support using UAVs equipped with cameras. However, a limitation of cameras is their difficulty in detecting individuals concealed in shadows. To solve this problem, we propose the use of a listening device on UAVs to detect sounds created by humans. This device uses an on-UAV microphone to capture human voices, which often get mixed with the sound of the UAV's propellers. This mixing presents a major challenge in identifying human voices. In this paper, we introduce a method to suppress the UAV propeller sound noise from the mix, enhancing the clarity of the human voice. Suppression of UAV sound noise is performed by generating pseudo-UAV sound based on generative adversarial networks (GAN) and reducing the generated pseudo-UAV sound from the sound mixture. By conducting various types of experiments, we confirmed the effectiveness of our proposal. As a result, we established the feasibility of using UAV-based voice processing for victim detection at disaster sites.**

***Index Terms*—Generative Adversarial Networks, Victim Detection, Rescue Support, UAV Application, Sound Noise Suppression**

I. INTRODUCTION

IN recent years, research has focused on utilizing unmanned aerial vehicles (UAVs) to assess damage and conduct search and rescue operations for victims during natural disasters [1-3]. Searching for victims in collapsed houses and rubble, particularly in the aftermath of earthquakes, necessitates early detection and rapid, accurate response. The use of UAVs in such scenarios is critically important and beneficial, as they facilitate easy access to areas otherwise difficult to reach due to the disaster [4-11]. However, most existing search and rescue operations using UAVs rely primarily on visual confirmation through images captured by cameras mounted on the UAVs [12]. While these approaches allow for assessing the situation captured in the video, it is challenging to recognize victims who are not visible, such as those trapped under rubble or in the camera's blind spots. One study sought to eliminate camera blind spots [13] by using two 360° cameras. However, these cameras pose challenges when mounted on UAVs, as the UAV itself is often captured in the images. Additionally, the spherical lens distorts the visual representation of objects, complicating the assessment of immobile victims[14]. A proposed solution for victim detection involves using a UAV-mounted voice processing system, which includes an onboard microphone and speaker. This system broadcasts calls into the disaster area and listens for responses using the UAV's microphone. When a response is detected, it indicates the presence of a victim [15]. However, the challenge lies in the sound captured by the microphone, which is a mix of the victim's voice, UAV propeller noise, and environmental sounds. Therefore, it's crucial to effectively isolate the victim's voice from this mixture for accurate detection. Currently, similar studies are underway into finding victims based on the voice obtained from microphones mounted on UAVs by using a system called source separation, in which multiple microphones are used to locate the source of a sound separately from the propeller noise emitted by the UAV [16-20]. In these systems, however, it is not possible to determine the content of the sound source, such as the words spoken by the victim, because it determines only the location of the sound source. In addition, all sounds, including the propeller noise, are processed in the same way regardless of whether they are human or not. Another method is to determine the words uttered by the victims using speech recognition software [18]. However, the challenge arises because the words used for speech recognition are predetermined, while the words spoken by a victim vary based on the situation. This makes it difficult to recognize phrases spoken by victims if they are not already registered in the speech recognition software. Furthermore, as the victim's voice is processed by the voice recognition software, the UAV operator does not hear it directly, preventing them from verifying the accuracy of the recognition. The accuracy of voice recognition software significantly decreases with the distance between the victim and the UAV, as illustrated in Figure 1 [18]. In time-critical situations, incorrect decisions can have dire consequences, especially for severely injured individuals. Therefore, it is recommended to first listen to a victim's voice for a more precise assessment of their condition. Listening to victims not only helps in accurately understanding their

This work was supported in part by the Japan Society for the Promotion of Science – Grant-in-Aid for Scientific Research (C) (Grant No. 21K04592).

Chinthaka Premachandra is a Professor with the Department of Electronic Engineering, School of Engineering/Graduate School of Engineering and Science, Shibaura Institute of Technology, Tokyo, Japan (e-mail: chintaka@shibaura-it.ac.jp).

Yugo Kunisada was a master student with the Department of Electronic Engineering, Graduate School of Engineering and Science, Shibaura Institute of Technology, Tokyo, Japan (e-mail: ma20035@shibaura-it.ac.jp))

situation but also aids in determining the necessary supplies and assistance required, particularly for those in hard-to-reach areas.

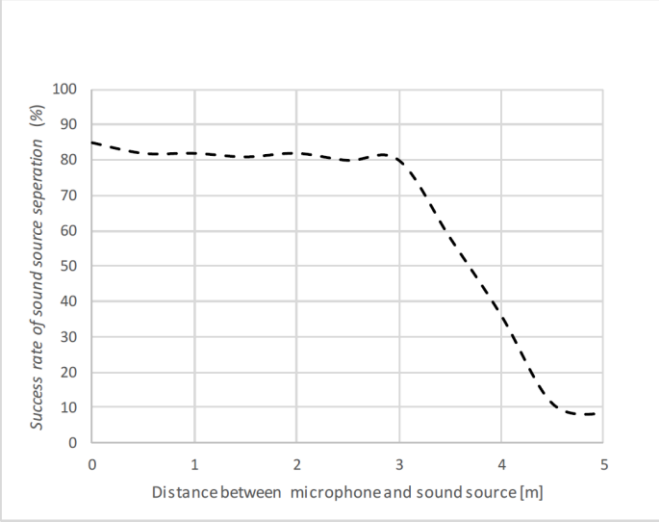


Fig. 1. Sound source separation results of the method proposed by Yamazaki et al [18].

This study aims to process sounds, including the UAV's propeller noise, to make the victim's voice audible to the UAV operator by suppressing the UAV sound noise from sound captured by UAV onboard microphone. This approach allows the operator to directly hear the victim's voice, rather than relying on indirect detection methods to ascertain the presence of a victim.

A schematic of the victim detection system developed in this study is illustrated in Fig. 2, while Fig. 3 depicts the procedure for the proposed method of suppressing UAV propeller sound. In this research, we utilize the advanced capabilities of generative adversarial networks (GANs) [23], a form of AI that has recently undergone significant improvements in accuracy, to learn various types of data, including images, frequencies and so on. Specifically, we utilize GANs to learn and analyze the propeller sound noise emitted by UAVs. Using the model trained in this manner, we are able to generate pseudo-UAV sound. By subtracting this pseudo-UAV sound from the actual sound captured by the UAV-mounted microphone, we effectively suppress sounds at the same frequency in the real UAV sound. Our method holds several advantages over traditional signal processing-based noise filters: (1) it more precisely suppresses UAV noise within a narrower bandwidth, (2) it adapts to fluctuating UAV noise via the learning model, and (3) it achieves higher accuracy in UAV noise suppression. The method's ability to handle fluctuating UAV noise (above (2)) is particularly noteworthy when compared with conventional methods. In this study, we apply the Fast Fourier Transform (FFT) [22] to the sound of UAV propellers as a preprocessing step for GAN-based machine learning, using frequency component as training data. This approach, which involves learning sound separately on each frequency, is distinct from common sound learning methods for speech recognition. To the best of our knowledge, this research is innovative in adapting sound learning to noise cancellation.

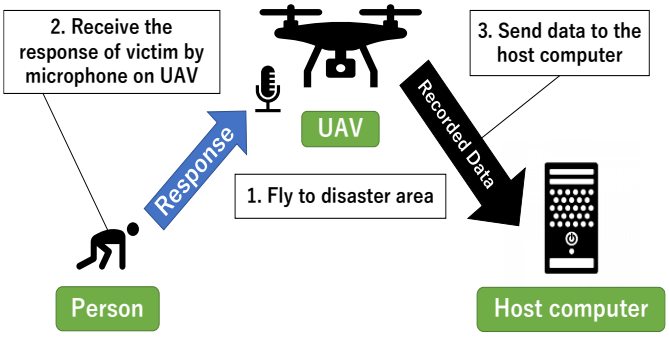


Fig. 2. A schematic of the victim detection system.

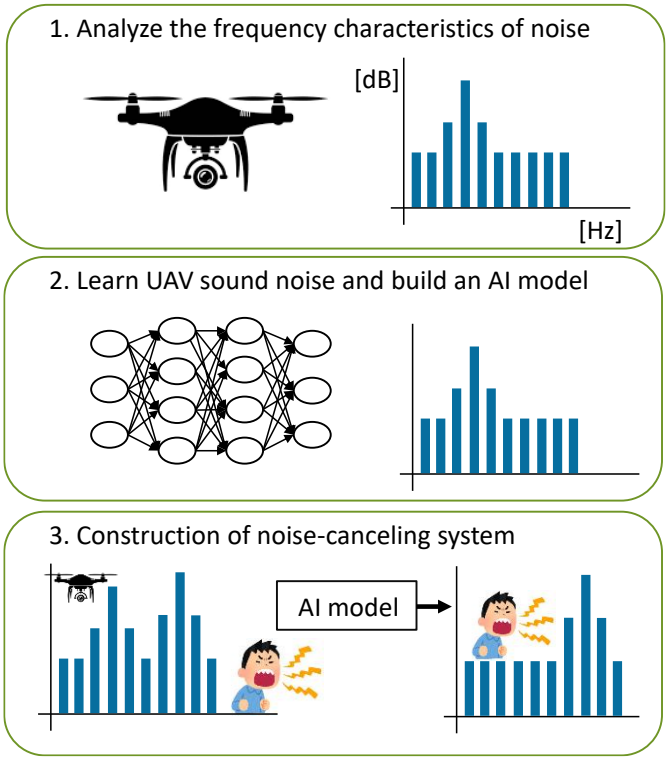


Fig. 3. Overall flow of human voice detection based on UAV propeller sound suppression.

Our approach involved several key steps: (1) constructing and training a Generative Adversarial Network (GAN) model with optimal preprocessing to generate pseudo-UAV sound, (2) assessing whether the pseudo-UAV sound produced by the model is effective for noise suppression, (3) testing the efficiency of the proposed noise suppression process with actual mixtures of UAV sound and human voice, (4) designing a UAV equipped with the necessary hardware for this study, and (5) developing a specialized off-board host service computer system, focused on voice learning and generating pseudo-UAV sound, as a critical hardware development in this study.

This paper is organized into five sections, offering a comprehensive overview of the entire project, encompassing both software and hardware developments. Section 2 details the UAVs used in the study, including their specifications, the functionalities of their onboard computer systems, and how these systems integrate with a developed off-board service host computer system. Section 3 describes the proposed UAV sound noise suppression method, focusing on learning theory and incorporating mathematical models and expressions related to Generative Adversarial Networks (GANs). Section 4 presents the experimental results, outlining the data collection, analysis, and implications. The paper concludes in Section 5 with a summary of the findings, a discussion of the study's limitations, and suggestions for future research.

II. Hardware Architecture

Fig. 4 displays the UAV developed for this research. We employed a Raspberry Pi and Navio2 as flight controllers for the UAV, chosen for their convenience in computer access and program development. The UAV is equipped with a microphone, essential for future research developments. This microphone features a quad array, although in this study, only the central microphone was utilized.

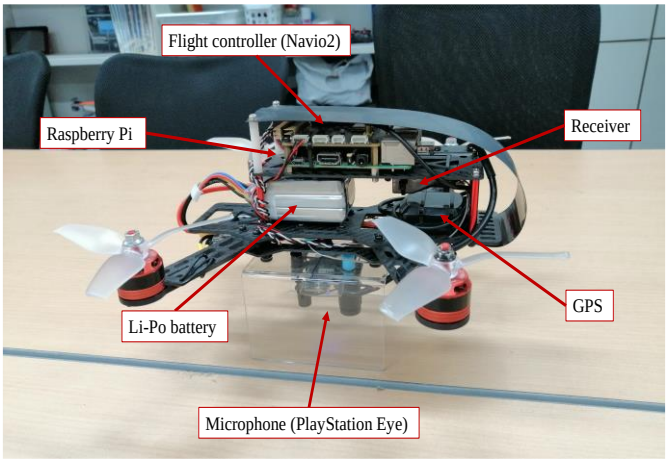


Fig. 4. Overall view of developed UAV.

In this study, we focus on sound learning and the generation of pseudo-UAV sounds, which are conducted on the host service computer. Fig. 5 presents an overall diagram of the computer system architecture made in our research. Initially, attempts were made to process the data using a computer mounted on the UAV. However, this approach was abandoned due to the limited processing power of the onboard processor. As a result, the computer mounted on the UAV was adapted to function primarily connecting to a Wi-Fi router. This setup enables the transmission of voice data, captured by the onboard microphone, to the host computer using the User Datagram Protocol (UDP). The audio data was recorded by the onboard microphone with a sampling frequency of 16 kHz and a bit depth of 16 bits. Excluding the flight controller, all processing tasks for both the UAV and the host computer were conducted using Python. To enhance GAN processing speed on the host computer, a GPU, specifically a NVidia GTX1660, was used in conjunction with TensorFlow and Keras for building the machine learning models. This system configuration aligns with the concept of transitioning UAVs to computers with powerful processors for edge computing.

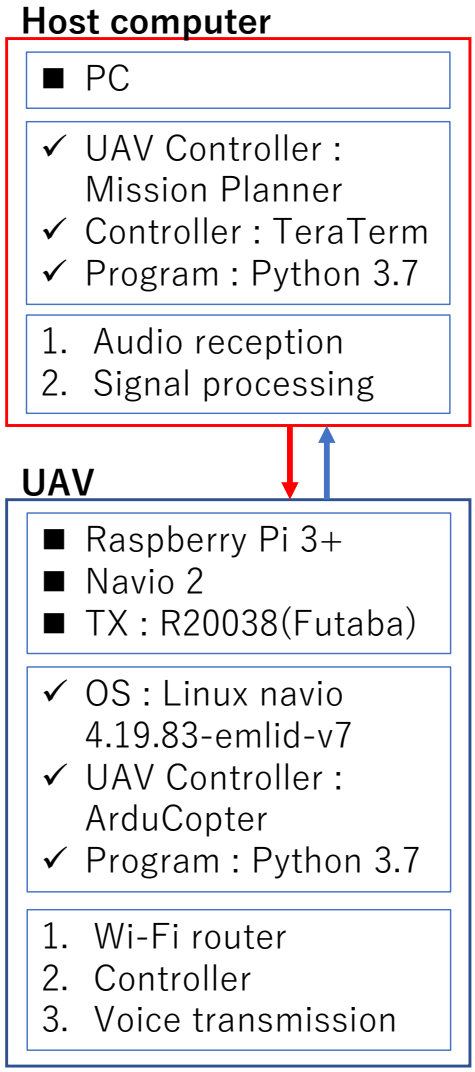


Fig. 5. Overview of the on-board hardware and off-board host computer architecture.

III. UAV Sound Suppression Process

In this study, we propose a noise cancellation technique specifically intended to suppress UAV sounds within audio recorded by the UAV's microphone. The method entails training a machine learning model using authentic UAV sounds and subsequently generating a pseudo-UAV sound from this model. This artificially created pseudo-UAV sound is then subtracted from the audio captured by the UAV's microphone, effectively cancelling (suppressing) the UAV sound noise (Fig. 6).The choice of this method is motivated by the characteristic frequency of UAV sounds.

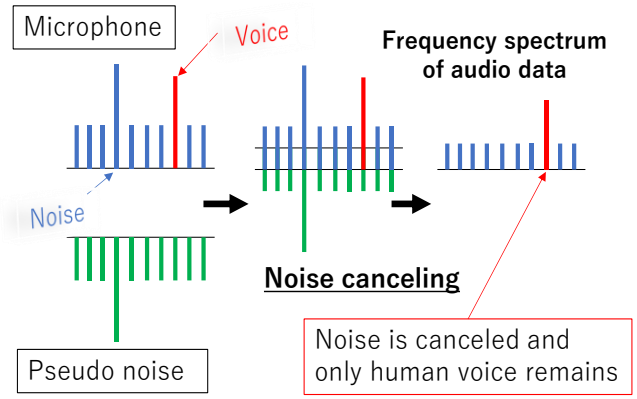
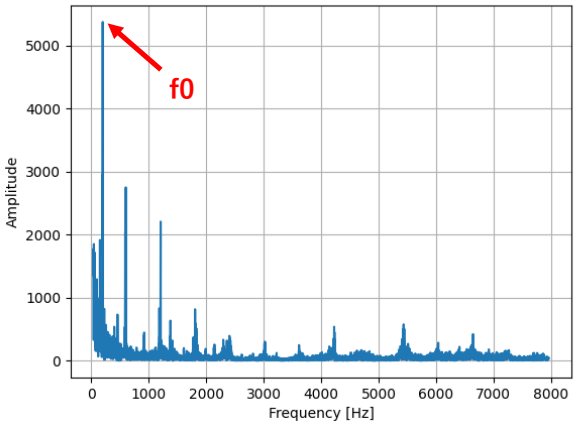


Fig. 6. Outline of the noise canceling (suppression) procedure of this study.

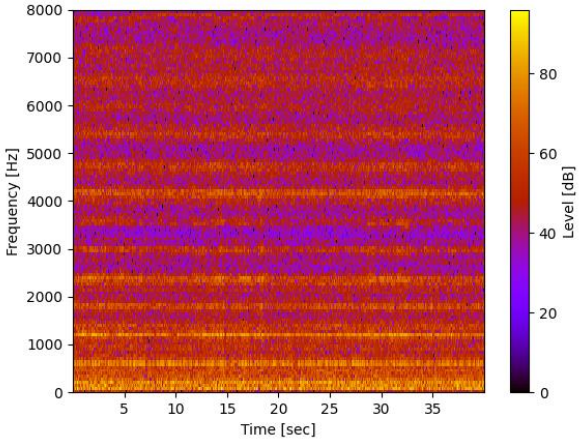
A. *UAV Sound Data Characteristic Analysis*

In this study, we observed that the sound produced by the UAV varies with each movement, including changes in altitude and forward or backward motion. The frequency band of the UAV sound tends to rise with increased motor output and fall when the output decreases. We recorded and analyzed audio data at different motor power levels, under the assumption that the audio characteristics change with actual UAV operation. The frequency spectrums of the UAV sound at 50%, 75%, and 100% motor power of a propeller are depicted in Figs. 7(a), 8(a), and 9(a) respectively. In these spectrum diagrams, the maximum amplitude of the audio data is indicated as f0. Additionally, spectrograms of propeller sound, showing time and frequency at each motor output level are presented in Figs. 7(b), 8(b), and 9(b). These spectrograms reveal that the frequency components of the UAV sound are relatively stable at a constant motor power. Specifically, from Fig. 7, the UAV sound at 50% motor power has a prominent component around 200 Hz. When compared with Figs. 8 and 9, which represent 75% and 100% motor power respectively, we observed that the UAV sound frequency and noise level increase significantly with higher power, more than doubling in some cases. However, the comparison between 75% and 100% motor power shows a frequency shift of about 50 Hz without a substantial change in sound volume level. This variability suggests that a simple noise filter, which cannot adapt to the fluctuating frequency range of a flying UAV's sound, is inadequate. Creating a system to automatically adjust the filter's cut-off frequency based on varying frequencies is not practical because it relies on audio levels and it's hard to accurately suppress UAV sounds. Consequently, noise filtering based on conventional signal processing is impractical for this application. Thus, based on these findings, we opted for a machine learning model that learns from data with values on the frequency axis for this study.



	[dB]	[Hz]
f0	74.6	201.0

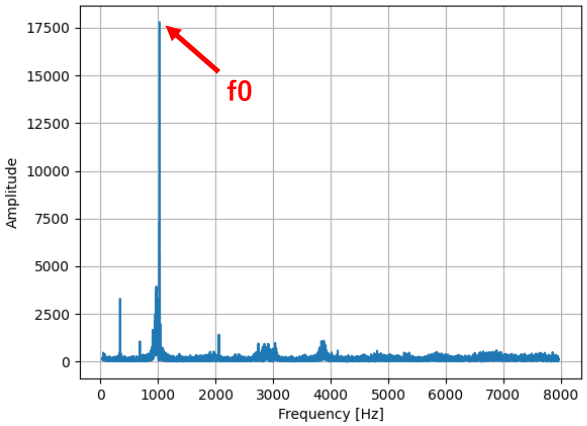
(a)



(b)

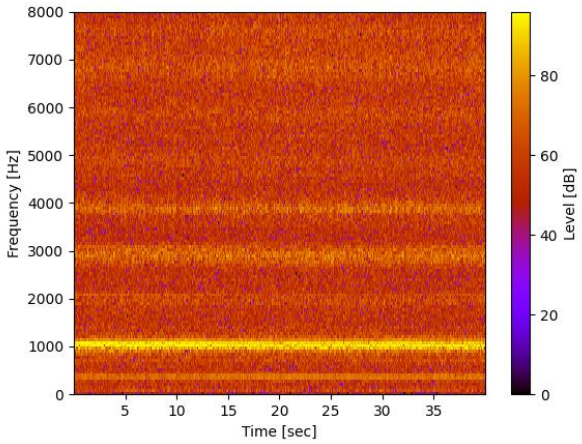
Fig. 7. Spectrum and spectrogram analysis of Rotor Motor Power at 50% (a) Spectrum of output motor power at 50%. (b) Spectrogram of output motor power at 50%.

The technique of employing frequency-dependent audio data for training in this study stands in contrast to the models typically used in speech recognition, which rely on the waveform of speech to identify words. In conventional audio learning for speech recognition, words need to be recognizable irrespective of the speaker's gender or tone, indicating that frequency alone is insufficient for learning. Our study doesn't concentrate on recognizing human speech, but rather on removing unwanted sounds from a mixture that includes human speech to improve the clarity of the speech. Additionally, the audio data we aim to train is a continuous standing wave, making the use of frequency-dependent data more suitable. Therefore, as a preprocessing step, both the training audio data captured from the UAV's microphone underwent a Fast Fourier Transform (FFT) process. This step converted the time-dependent data into frequency-dependent data, preparing it for the subsequent training phase.



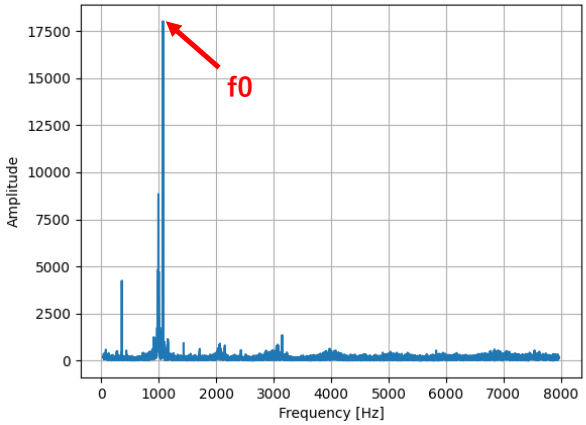
	[dB]	[Hz]
f0	85.0	1.03k

(a)

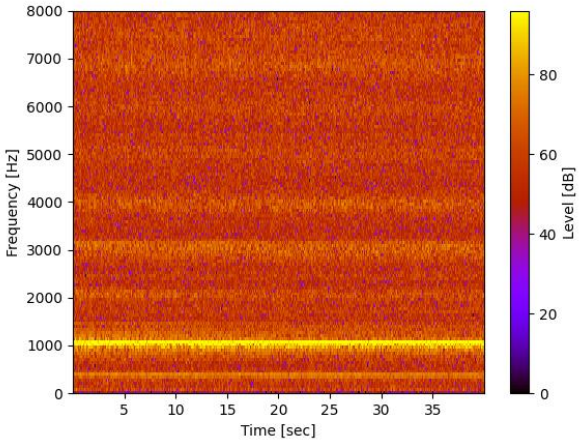


(b)

Fig. 8. Spectrum and spectrogram analysis of Rotor Motor Power at 75% (a) Spectrum of output motor power at 75%. (b) Spectrogram of output motor power at 75%.



	[dB]	[Hz]
f0	85.1	1.08k



(b)

Fig. 9. Spectrum and spectrogram analysis of Rotor Motor Power at 100% (a) Spectrum of output motor power at 100%. (b) Spectrogram of output motor power at 100%.

B. Application of GAN for Pseudo-UAV Sound Generation

This study employs a method known as Conditional GAN (CGAN) [24] to learn the sound of UAVs. CGAN, an evolution of the GAN method, has been applied in various learning contexts [25-27]. In our initial trials using only GAN, the output results for pseudo-UAV sound were found to be unsatisfactory, particularly in terms of frequency dependence. Consequently, we adopted CGAN to introduce specific conditions into the training data. Theoretically, since CGAN is derived from GAN, its basic structure mirrors that of GAN. Therefore, this section first introduces GAN and its learning methodology.

Typically, the GAN learning method has been used for image generation and has not been much used for audio frequency learning. In this study, we use the FFT of audio data as a preprocessing step for learning, and train each frequency as training data. A schematic diagram of the GAN training procedure is shown in Fig. 10. The GAN consists of two neural networks (NNs): a generator NN model (referred to as the Generator) and a discriminator NN (referred to as the Discriminator). In a fully trained GAN, the Generator becomes proficient enough at creating fake data that the Discriminator struggles to distinguish between real and fake inputs. In this paper, we train the Generator using the CGAN method, which is detailed below, to produce pseudo-UAV sound noise that closely adheres to the GAN. Subsequently, the trained Generator is utilized to generate pseudo-UAV sound data. The objective function of the GAN can be expressed in the form of Eq. (1).

$$min_G max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_z(x)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (1)$$

In the context of the audio data used in this study, $E_{x\sim p_{data}(x)}$ represents the expected value for the real UAV audio data used in training, and $D(x)$ signifies the Discriminator's evaluation of this real data. On the other hand,

$E_{x\sim p_{z(x)}}$ denotes the expected value when pseudo-UAV sound, which is synthetically generated by the Generator, is input to the Discriminator. Here, $G(z)$ refers to the pseudo-UAV sound created by the Generator from a noise input z, and $D\big(G(z)\big)$ indicates the Discriminator's evaluation of this generated sound. Essentially, while $D(x)$ assesses real audio data, $D\big(G(z)\big)$ assesses the authenticity of the pseudo-UAV sound produced by the Generator.

The entire function of eq. (1) forms a min-max game where the Generator tries to minimize the function (hence the min_G part) while the Discriminator tries to maximize it (hence the max_D part). This adversarial process leads to the Generator improving its ability to create data that resembles the real data, and the Discriminator improving its ability to differentiate real from fake.

In GAN, a one-dimensional uniform random value is input to the input layer of the generator, the output and the training data are mutually input to the discriminator, and the correctness is judged from the error. The weights of the Generator and Discriminator are modified based on the losses incurred. This learning technique enhances accuracy through repeated iterations and is primarily applied to discover correlations in large datasets of image data and to generate AI-created images [28-29].

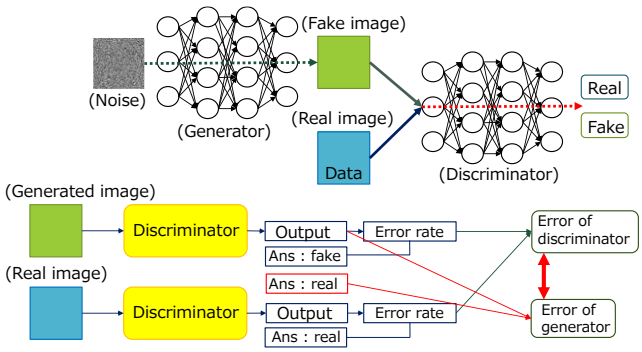


Fig. 10. Architecture of generative adversarial networks.

One of the advantages of using GAN in this context is that it does not require a large amount of supervised data across various output scenarios, which is typically necessary for enhancing the accuracy of other machine learning procedures. Traditional decision-making processes in machine learning, such as regression and classification, require a substantial and balanced amount of data for both positive and negative outcomes. However, with GAN, only the target data (positive data) for training is needed. This makes it particularly suitable for training with the specific audio data output by each motor of the UAV.

In this study, the training data have only two dimensions: frequency and the value of the level at that time. In addition, correlations can easily be determined by the characteristic UAV sound. Therefore, the use of a GAN in this study is effective both in terms of the amount training data and training time required. However, in order to learn the variation of UAV sound output by motor according to the varying flight status, we built a model partially based on a method called image-to-image translation with CGAN [30]. In the CGAN method, the input layer of the Generator or the input data can be pre-conditioned to produce arbitrary output results depending on the condition. The objective function of CGAN is given by Eq. (2) below.

$$min_G max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x,c)] + E_{x\sim p_z(x)}\left[\log\left(1-D\big(G(z,c)\big)\right)\right] \quad (2)$$

In this approach, diverging from the standard GAN as described in Eq. (1), a condition c is incorporated into the training data. This modification results in both the Generator and Discriminator being trained with the influence of condition c, thereby shaping the training process in specific ways. Other symbols in eq. (2) are the same as in Eq. (1).The primary aim of this study is to enable the Generator to produce outputs that closely match the ground truth. In alignment with this objective, the Generator's loss function is articulated in Eq. (3). $G(x)$ denotes the output of the Generator when it is given noise x as input while $D\big(G(x)\big)$denotes the output of the Discriminator when given $G(x)$, the output of the Generator.

$$L(G) = E_{x\sim p_z(x)}\left[\log\left(x-D\big(G(x)\big)\right)\right] \quad (3)$$

To ensure that the outputs from the Generator closely resemble ground-truth values, the model in this study is trained using Generator outputs that mirror the conditions of the input values. In the realm of image-to-image translation with CGAN, a generator model incorporating L1 regularization is often constructed [30]. However, in this study, such a process is deemed unnecessary due to the lower dimensionality of the data that used. Regarding the generation of pseudo-UAV sound, the input to the Generator is actual sound recordings from microphones mounted on the UAV. This approach allows the pseudo-UAV sound to be generated based on varying UAV motor outputs, obviating the need for selecting a specific pseudo-sound for each specific scenario. Furthermore, this method reduces computer processing requirements as it eliminates the necessity for a distinct generator model for each case.

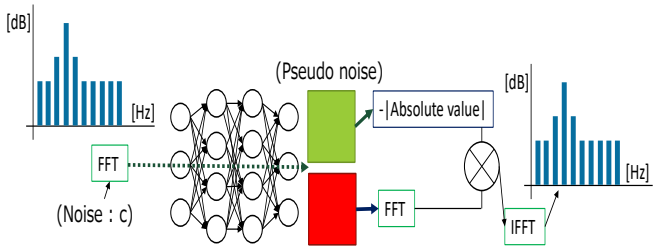


Fig. 11. CGAN model used in this study.

C. *UAV Sound Data Pre-processing for GAN*

Fig. 11 presents an outline of the CGAN utilized in this study.

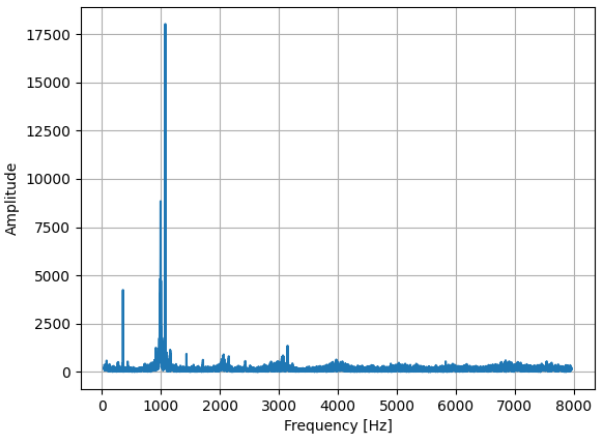
As a preliminary step, we simplify the data during pre-processing to enhance the learning of audio data correlations more effectively and rapidly. This simplification occurs prior to the learning phase, as depicted in Fig. 11. Neural Network (NN) training models in machine learning usually require input values ranging between 0 and 1, necessitating the normalization of the audio data obtained from the microphone. This normalization involves taking the absolute value of the audio data post-Fast Fourier Transform (FFT), a process outlined in Eq. (4). Here, each value of $F(t)$ corresponds to a particular frequency component of the original sound signal. $f(x)$ is the original signal in the time domain. A summation from $x = 0$ to $N - 1$, where N is the total number of samples in the signal. The summation is used to calculate each frequency component of the sound signal.

$$F(t) = \left| \sum_{x=0}^{N-1} f(x) exp\left(-i\frac{2\pi t x}{N}\right) \right| \quad (4)$$

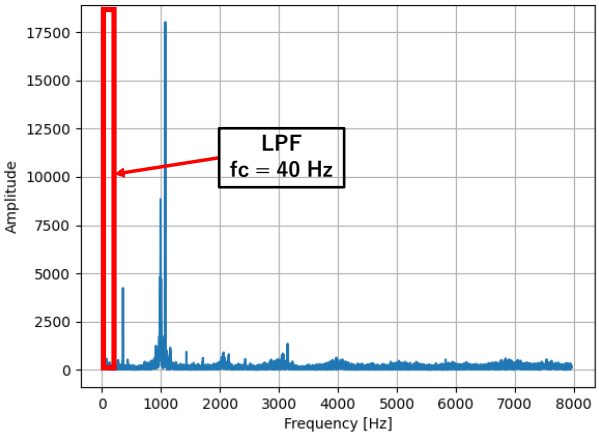
Although $F(t)$ in Eq. (4) has negative audio levels and negative frequencies, learning with NNs does not allow inputting negative values. In order to avoid this problem, we first biased all the audio level data so that the audio levels were positive. However, this method did not produce the desired learning results because the learning model could not capture the features. Therefore, the absolute values of all values were used as input values during training. In this study, the 16-bit depth digital audio data used has a maximum integer value of 65535. This value represents the highest possible amplitude in a 16-bit audio format. If normalization is performed by simply dividing the audio data by the maximum value of 65535, it results in excessively small values. This scale reduction makes it challenging to distinguish between different frequency components, which in turn causes a significant decrease in learning accuracy. In addition, when regenerating sound from the subtraction output values which is explained in next section, a challenge arises because multiplying the 16-bit values results in a significant deviation from the expected values, due to the small differences between each element's values. To address above issues, a preprocessing step was employed in this study. This step involves keeping the values of the normalized UAV sound data over 70 dB. Simultaneously, all irrelevant components in the audio data are set to 0 to focus the learning process on pertinent elements. This preprocessing approach is detailed in Eq (5), where $F(t) = X_t$.

$$\begin{matrix} X_t \leq 70[\text{dB}], & X_t = 0 \\ X_t > 70[dB], & X_t \end{matrix} \quad (5)$$

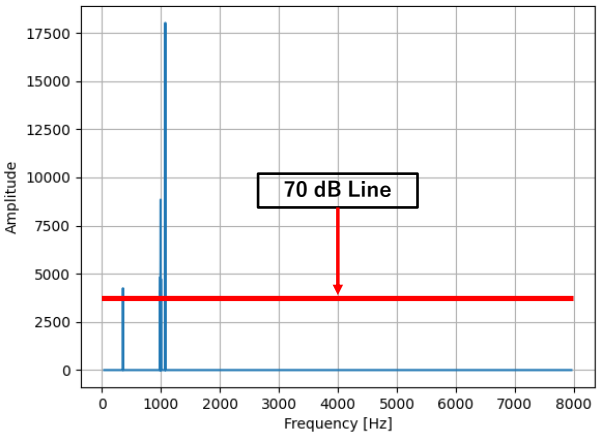
Furthermore, a low-pass filter (LPF) with a cut-off frequency of fc = 40 Hz was used in the low-frequency region to remove vibration noise from the UAV flight, added inrush noise due to microphone characteristics, and hum noise from the inverter. Fig. 12 shows the (a) input waveform, (b) LPF processing, and (c) processed waveform.



(a)



(b)



(c)

Fig. 12. Data preprocessing: (a) Original data; (b) LPF process; (c) Keeping the values of the normalized UAV sound data over 70 dB.

D. UAV Sound Noise Suppression

The objective of using a CGAN-trained generator model to generate pseudo-UAV sound is to process audio captured by a

microphone on an actual UAV. This pseudo-UAV sound is subtracted from the original audio recorded by the same UAV-mounted microphone. The purpose of this subtraction is to suppress the UAV noise, thereby enhancing the clarity of any human voices present in the recording. This technique is particularly beneficial when the UAV microphone captures audio mixed with human voices. The subtraction process is detailed in Eq. (6). To elaborate, the absolute value of the sound level data for each frequency, captured per second from the UAV's microphone, is represented as $|X_n| = [X_0, X_1, X_2, \dots]$. Similarly, the absolute value of the sound level data for each frequency per second for the pseudo-UAV sound, generated by the generator model, is represented as $|z_n| = [z_0, z_1, z_2, \dots]$. The result of subtracting $|z_n|$ from $|X_n|$ is denoted by $\omega_n = [\omega_0, \omega_1, \omega_2, \dots]$, where $n = [0,1,2, \dots, t/2]$.

$$\omega_n = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} X_0 \\ X_1 \\ \vdots \end{bmatrix} - \begin{bmatrix} z_0 \\ z_1 \\ \vdots \end{bmatrix} \quad (6)$$

In the process of manipulating audio data, the values for each frequency in ω_n may sometimes become negative. This occurs when the pseudo-UAV sound generated by the neural network is significantly louder than the actual input audio, resulting in an overshoot that is manifested as unwanted sound. Such negative values are contrary to the objectives of this research, which aims to isolate and enhance desired sounds like human speech. To rectify this issue, all negative values in ω_n are replaced with zero. This approach effectively prevents the reproduction of unwanted sound when the audio data is restored. The formal representation of this corrective measure is detailed in Eq. (7).

$$\omega_n \leq 0, \qquad \omega_n = 0 \quad (7)$$

E. Sound Restoration Processing

After suppressing noise (UAV sound) from the frequency data, the next step is to recover the audio data. Initially, the time-based audio data were transformed into frequency-based data using the Fast Fourier Transform (FFT). To revert these frequency-based data back into the time domain, the inverse FFT (IFFT) is employed. This transformation is crucial for reconstructing the original audio characteristics while excluding the noise components that were identified and suppressed earlier. Eq. (8) below illustrates the IFFT process, representing the discrete FFT as $F(t)$. Here, $Z(t)$ denotes the complex frequency domain representation of the signal obtained from the subtraction result. Other symbols are the same as in Eq. (4).

$$f(x) = \frac{1}{N}\sum_{t=0}^{N-1} Z(t) \exp\left(i\frac{2\pi t x}{N}\right) = \overline{\frac{1}{N}\sum_{t=0}^{N-1} \overline{Z(t)} \exp\left(-i\frac{2\pi t x}{N}\right)} \quad (8)$$

As shown in Eq. (6), since the value of ω_n is taken as the absolute value during the learning process, the positive and negative symbols of the original audio data are lost. In this case, sound restoration by IFFT is difficult. In the process of sound recovery, it's crucial to maintain the correct symbol (positive or negative) for each audio data point. To overcome this hardle, in our study, we employed the U-Net technique [32][33], as shown in Fig. 13. The U-Net method enables the reintegration of certain information at both the input and output stages, circumventing the learning phase of the GAN's neural network. This feature is crucial for preventing the loss of significant data during the input process.

As depicted in Fig. 14, the U-Net system in this study plays a key role in ensuring the integrity of the symbol, phase, and volume of data. Notably, due to the linear symmetry of the frequency values around axis 0 in the FFT, we trained only the values on the positive axis of the frequency spectrum. The results were then linearly replicated on the negative axis. This approach simplifies the process while preserving the accuracy of the audio data. As a result, it is possible to use only half of the total training data, which greatly reduces the training time and processing requirements of the PC used for training.

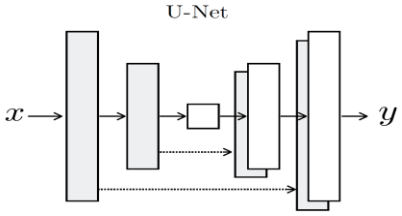


Fig. 13. The basic U-Net architecture.

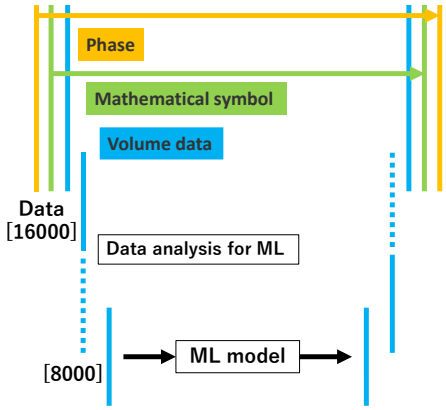


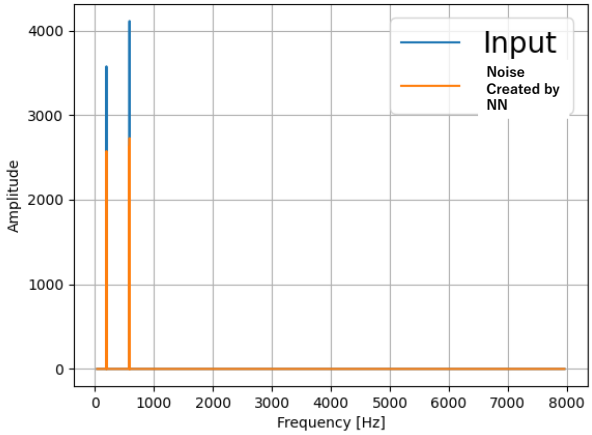
Fig. 14. The U-Net application of this research.

IV. Experimental Evaluation

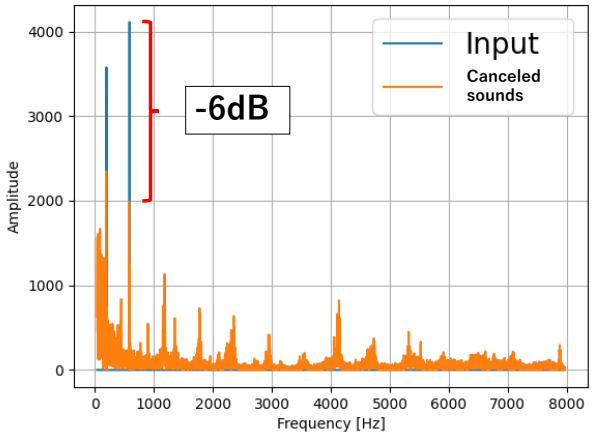
A. Experimental Environment

To verify the effectiveness of the proposed sound suppression method, a learning model was used to learn the sounds of UAVs and to generate pseudo-UAV sounds at 50%, 75%, and 100% motor power levels of the rotors. The training audio data have a sampling frequency of 16 kHz and a bit depth of 16 bits. During the experiments, the UAV was positioned approximately 3 meters away from the human subject. This distance was considered sufficient for effectively recording the human voice, considering the sensitivity of the microphone

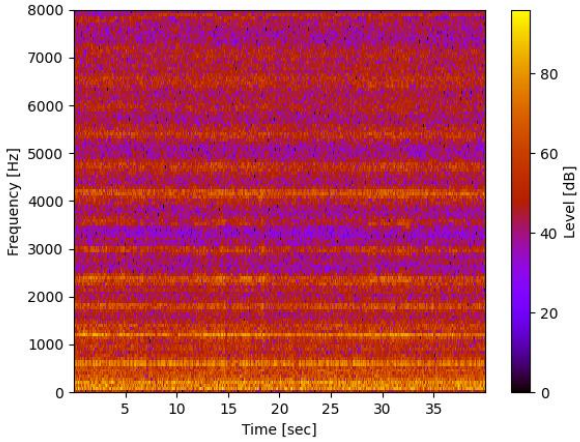
used. The range can be extended by employing a more sensitive microphone. The audio data consist of recordings of the UAV sound at each motor power level, with 5 minutes of audio data used for each training session. Audio data containing a mixture of UAV sounds and human voices were also prepared in the same manner. Consequently, the model was trained on a dataset that included data for each frequency, comprising a total of 300 data points.



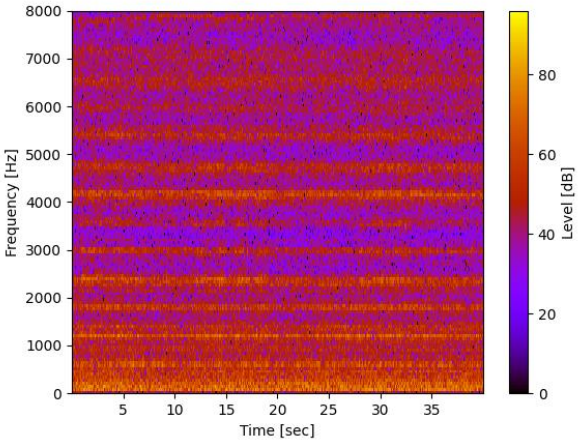
(a)



(b)

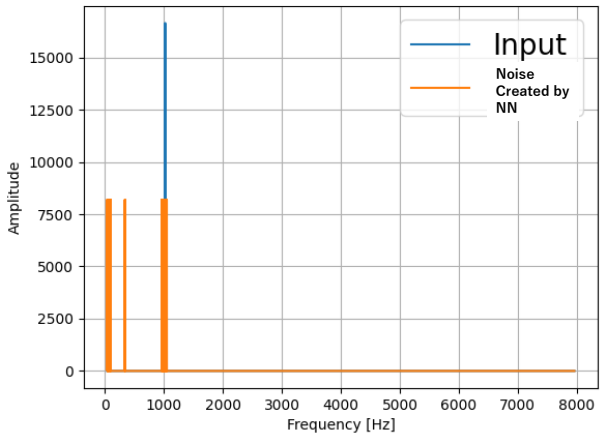


(c)

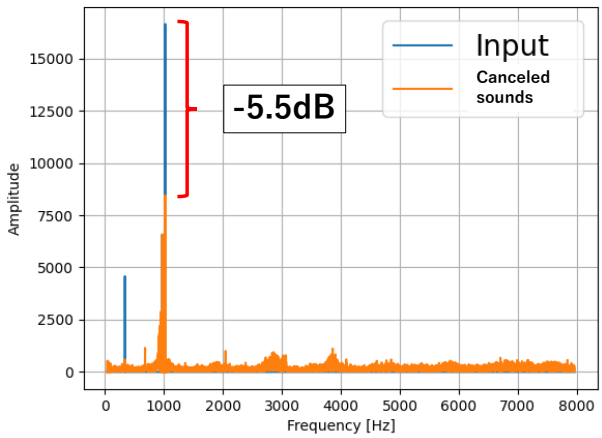


(d)

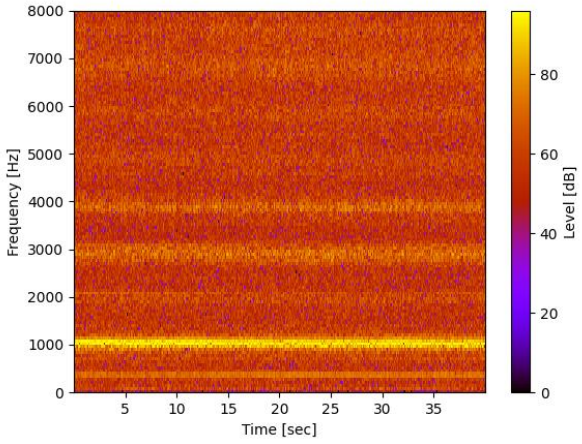
Fig. 15. Cancelling (suppression) of UAV sound noise at rotor motor power 50%. (a) the input audio frequency spectrum and a comparison with the frequency spectrum of the pseudo-UAV sound generated by GAN Generator model, (b) the frequency spectrum after subtracting the UAV sound, along with a comparison with the input audio frequency spectrum, (c) the input audio frequency spectrogram, and (d) the frequency spectrogram after cancelling (suppressing) the UAV sound.



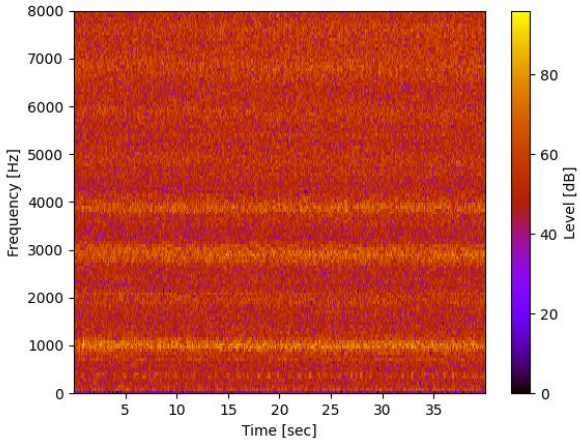
(a)



(b)

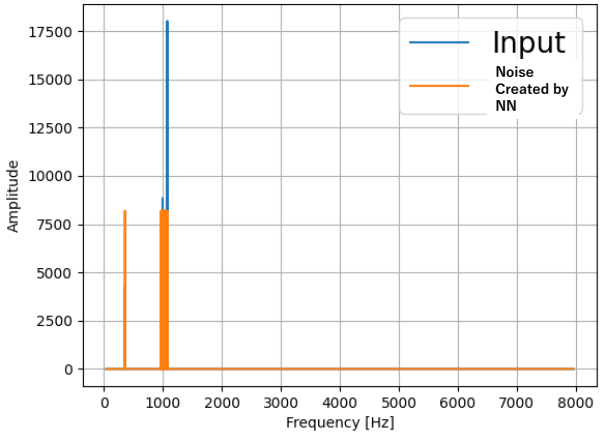


(c)

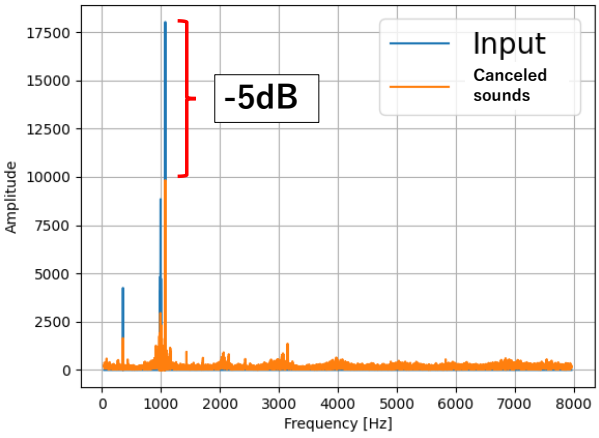


(d)

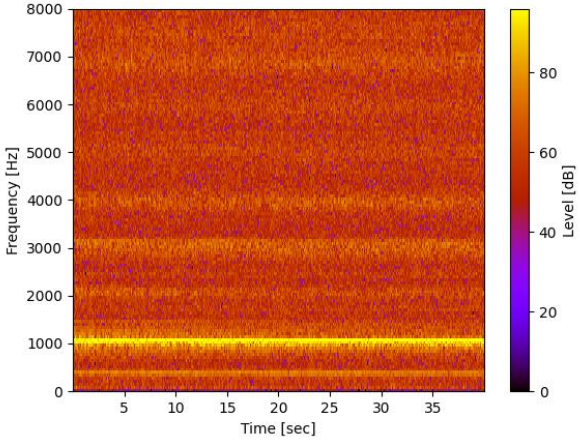
Fig. 16. Cancelling (suppression) of UAV sound noise at rotor motor power 75%. (a) the input audio frequency spectrum and a comparison with the frequency spectrum of the pseudo-UAV sound generated by the GAN Generator model, (b) the frequency spectrum after subtracting the UAV sound, along with a comparison with the input audio frequency spectrum, (c) the input audio frequency spectrogram, and (d) the frequency spectrogram after cancelling (suppressing) the UAV sound.



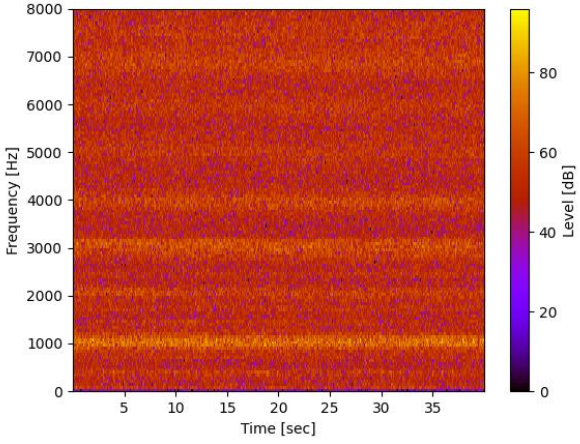
(a)



(b)



(c)



(d)

Fig. 17. Cancelling (suppression) of UAV sound noise at rotor motor power 100%. (a) the input audio frequency spectrum and a comparison with the frequency spectrum of the pseudo-UAV sound generated by the GAN Generator model, (b) the frequency spectrum after subtracting the UAV sound, along with a comparison with the input audio frequency spectrum, (c) the input audio frequency spectrogram, and (d) the frequency spectrogram after cancelling (suppressing) the UAV sound.

B. *Experimental Results*

In this paper, we primarily focus on evaluating the suppression of UAV sound. Figures 15, 16, and 17 illustrate: (a) the input audio frequency spectrum and a comparison with the frequency spectrum of the pseudo-UAV sound generated by the proposed Generator model based on GAN, (b) the frequency spectrum after subtracting the UAV sound, along with a comparison with the input audio frequency spectrum, (c) the input audio frequency spectrogram, and (d) the frequency spectrogram after cancelling (suppressing) the UAV sound, using trained generator models at 50%, 75%, and 100% UAV rotor motor power, respectively.

Comparing Figures 15(a), 16(a), and 17(a), we observe that the pseudo-UAV sound is generated along the frequency axis, corresponding to the UAV sound component output by each motor. Figures 15(b), 16(b), and 17(b) demonstrate that the UAV noise-suppression process effectively reduces the UAV noise component by an average of -5 dB across all outputs.

These findings indicate that the machine learning method employed for generating pseudo-UAV sound can accurately learn the frequency and amplitude levels of the sound noise components corresponding to motor's output. The results depicted in the frequency-amplitude level graphs reveal that the UAV sound noise component is reduced by approximately 5 dB on average, a reduction deemed appropriate and effective for noise suppression. Upon listening to the restored audio data, we perceived a noticeable suppression in the UAV noise level. However, it should be noted that this method does not eliminate all UAV sound components.

The sound regeneration process was also tested by generating the remaining sound after UAV sound noise suppression, particularly focusing on the sound mixture of UAV sound and human voice. This process achieved success to a certain degree; however, some noise persisted in the remaining audio data. Despite this, the audibility of the human voice was notably improved. This improvement will enable the application of the proposed UAV sound suppression for victim detection activities at disaster sites, using on-board UAV microphones. The results of these tests can be verified through the video attached to this paper.

All the aforementioned experiments were conducted using a specially developed hardware architecture. This setup involved connecting a UAV onboard type small computer with the offboard host service computer. Throughout these experiments, we also verified the implementation capabilities of this hardware environment.

V. Conclusion

In this study, we introduced a novel audio processing method and demonstrated its effectiveness in overcoming the challenges faced by voice-based systems when utilizing Unmanned Aerial Vehicles (UAVs) in disaster scenarios. Initially, we developed the necessary hardware architecture, establishing a connection between the UAV's onboard small-type computer and an offboard service host computer. Our approach focused on suppressing the sound generated by UAVs from a mixture of UAV and human voice sounds, with the aim of enhancing the clarity and audibility of the human voice. This suppression process entails generating UAV sound using a Generative Adversarial Network (GAN) and then subtracting this generated sound from the mixed audio. Additionally, we present a method for regenerating the residual sound post-subtraction, employing a U-net architecture.

In our experiments, pseudo-UAV sounds at varying UAV rotor motor powers were generated and subtracted from actual audio data comprising both UAV and human voice sounds, to assess the system's efficacy. Our method proved to be somewhat effective in amplifying the human voice when mixed with UAV sounds. The current performance would be enough to apply the proposal for human detection process at disaster sites. However, it was observed that some noise still remained in the resultant audio. Based on these findings, we plan to continue our research, focusing on underlying issues that have emerged. Ultimately, our goal is to implement GAN on low-end edge computers [34], enhancing real-time processing capabilities.

References

[1] G. -J. M. Kruijff et al., "Rescue robots at earthquake-hit Mirandola, Italy: A field report," 2012 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), 2012, pp. 1-8.

[2] A. Albanese, V. Sciancalepore, X. Costa-Perez, "SSARDO: An Automated Search-and-Rescue Drone-based Solution for Victims Localization," IEEE Transactions on Mobile Computing, 2021.

[3] M. Erdelj and E. Natalizio, "UAV-assisted disaster management: Applications and open issues," 2016 International Conference on Computing, Networking and Communications (ICNC), 2016, pp. 1-5.

[4] Y. Ham, "Visual monitoring of civil infrastructure systems via cameraequipped Unmanned Aerial Vehicles (UAVs): a review of related works," Visualization in Engineering, 2016.

[5] S. Lee, D. Har and D. Kum, "Drone-Assisted Disaster Management: Finding Victims via Infrared Camera and Lidar Sensor Fusion," 2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2016, pp. 84-89.

[6] A. Gaszczak, T. P. Breckon, and J. Han, "Real-time People and Vehicle Detection from UAV Imagery," Proc. of the SPIE, Vol. 7878, Jan 2011.

[7] S. Yang, X. Yang, and J. Mo, "The application of unmanned aircraft systems to plant protection in china," Precis. Agric., Vol. 19 (2), pp. 278-292, 2018.

[8] X. Li, Y. Zhao, J. Zhang and Y. Dong, "A Hybrid PSO Algorithm Based Flight Path Optimization for Multiple Agricultural UAVs," 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), 2016, pp. 691-697.

[9] D. Mellinger and V. Kumar, "Minimum snap trajectory generation and control for quadrotors," 2011 IEEE International Conference on Robotics and Automation, 2011, pp. 2520-2525.

[10] J. Yoon, I. Kim, W. Chung and D. Kim, "Fast and accurate car detection in drone-view," 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), 2016, pp. 1-3.

[11] Y. Karunarathna, H. Kawanaka, C. Premachandra and S. Tsuruoka, "Eagle Eye for ER Doctor: Basic Study on Drone based TeleMonitoring System for an Inaccessible Area," 2020 International Conference on Image Processing and Robotics (ICIPRoB), 2020, pp. 1-5.

[12] M.A.R. Estradaa, A. Ndomab, "The uses of unmanned aerial vehicles – UAV's- (or drones) in social logistic: Natural disasters response and humanitarian relief aid," ICTE in Transportation and Logistics 2018 (ICTE 2018), 2019, pp. 375-383.

[13] M. Tamaki and C. Premachandra, "An Automatic Compensation System for Unclear Area in 360-degree Images Using Pan-tilt Camera," 2019 International Symposium on Systems Engineering (ISSE), 2019, pp. 1-4.

[14] C. Premachandra, S. Ueda and Y. Suzuki, "Detection and Tracking of Moving Objects at Road Intersections Using a 360-Degree Camera for Driver Assistance and Automated Driving," in IEEE Access, vol. 8, pp.

135652-135660, 2020.
[15] Y. Yamazaki, M. Tamaki, C. Premachandra, C. J. Perera, S. Sumathipala and B. H. Sudantha, "Victim Detection Using UAV with On-board Voice Recognition System," 2019 Third IEEE International Conference on Robotic Computing (IRC), 2019, pp. 555-559.
[16] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa and H. Tsujino, "An open-source software system for robot audition HARK and its evaluation," Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots, 2008, pp. 561-566.
[17] K. Nakadai et al., "Development of microphone-array-embedded UAV for search and rescue task," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 5985-5990.
[18] Y. Yamazaki, C. Premachandra and C. J. Perea, "Audio-Processing-Based Human Detection at Disaster Sites with Unmanned Aerial Vehicle," in IEEE Access, vol. 8, pp. 101398-101405, 2020.
[19] I. V. McLoughlin, "Super-Audible Voice Activity Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 9, pp. 1424-1433, Sept. 2014.
[20] D. E. Badawy, N. Q. K. Duong and A. Ozerov, "On-the-Fly Audio Source Separation—A Novel User-Friendly Framework," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 2, pp. 261-272, Feb. 2017.
[21] A. Magassouba, N. Bertin and F. Chaumette, "Aural Servo: Sensor-Based Control From Robot Audition," in IEEE Transactions on Robotics, vol. 34, no. 3, pp. 572-585, June 2018.
[22] J. W. Cooley, J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," Math. Comput. 19, 1965, pp. 297–301.
[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative adversarial nets,” Proc. NIPS 2014, pp. 2672-2680.
[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets, " arXiv preprint arXiv:1411.1784, 2014.
[25] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396-4405.
[26] D. Kim, J. Cha, S. Oh and J. Jeong, "AnoGAN-Based Anomaly Filtering for Intelligent Edge Device in Smart Factory," 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), 2021, pp. 1-6.
[27] M. Önder and Y. S. Akgül, "Automatic Generation of Matching Clothes Design Using Generative Adversarial Networks," 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1-4.
[28] A. Radford, L. Metz, S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.
[29] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242-2251.
[30] P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967-5976.
[31] R. Nakagomi, S. Arai, K. Furihata, "Measurements of acceptable levels, unacceptable levels and uncomfortable loudness levels of environmental sounds in persons with normal hearing," Audiology Japan 54(2), 2011, pp. 138-146.
[32] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention (MICCAI),2015, pp. 234-241.
[33] Y. Kunisada and C. Premachandra, "Sound-to-Sound Translation Using Generative Adversarial Network and Sound U-Net," Proc. of 2nd International Conference on Image Processing and Robotics, March 2022.
[34] A. S. Winoto, M. Kristianus and C. Premachandra, "Small and Slim Deep Convolutional Neural Network for Mobile Device," in IEEE Access, vol. 8, pp. 125210-125222, 2020.

Chinthaka Premachandra (Senior Member, IEEE) was born in Sri Lanka. He received the B.Sc. and M.Sc. degrees from Mie University, Tsu, Japan, in 2006 and 2008, respectively, and the Ph.D. degree from Nagoya University, Nagoya, Japan, in 2011. From 2012 to 2015, he was an Assistant Professor with the Department of Electrical Engineering, Faculty of Engineering, Tokyo University of Science, Tokyo, Japan. From 2016 to 2017, he was an Assistant Professor with the Department of Electronic Engineering, School of Engineering, Shibaura Institute of Technology, Tokyo, where he was an Associate Professor, from 2018 to 2022. In 2022, he was promoted to a Professor with the Department of Electronic Engineering, Graduate School of Engineering, Shibaura Institute of Technology, where he is currently the Manager of the Image Processing and Robotic Laboratory. His research interests include AI, UAV, image processing, audio processing, intelligent transport systems (ITS), and mobile robotics.

He is a member of IEEE, IEICE, Japan; SICE, Japan; RSJ, Japan; and SOFT, Japan. He received the IEEE SENSORS LETTERS Best Paper Award from the IEEE Sensors Council in 2022 and the IEEE Japan Medal from the IEEE Tokyo Section in 2022. He also received the FIT Best Paper Award and the FIT Young Researchers Award from IEICE and IPSJ, Japan, in 2009 and 2010, respectively. He has served as a steering committee member and an editor for many international conferences and journals. He is the Founding Chair of the International Conference on Image Processing and Robotics (ICIPRoB), which is technically co-sponsored by IEEE. He is currently serving as an Associate Editor for IEEE Robotics and Automation Letters (R-AL) and IEICE Transactions on Information and Systems.

Yugo Kunisada received B.S degree in electronic engineering from Shibaura Institute of Technology,Tokyo, Japan, in 2020. He received his M.S degree in Graduate School of Engineering and Science, Shibaura Institute of Technology, Tokyo, Japan, in 2022.

His research interests include Audio processing, pattern recognition, and human support systems.