Synesthetic Information Collecting Method by Auditory Visualization & Multivariate linear regression

Gyumin Cho
Gwangju Institute of
Science and Technology
123, Cheomdangwagi-ro,
Buk-gu, Gwangju,
Republic of Korea
South Korea
chocumin@gm.gist.ac.kr

Chang Wook Ahn*
Gwangju Institute of
Science and Technology
123, Cheomdangwagi-ro,
Buk-gu, Gwangju,
Republic of Korea
South Korea
cwan@gist.ac.kr

ABSTRACT

Humans collect auditory information through the sense of the eardrum and visual information through the sensory organ of the eye. In the case of sensory deficiency or damage, sensory information cannot be collected correctly. Alternatively, to solve these problems, we propose a synesthetic information collecting method in this paper. This method converts auditory information into visual information by taking the video of the vibrating object vibrating is used as the visual information for this paper. In the preprocessing process for extracting auditory information from this visual information, phase-based video magnification could be used to magnify minute vibrations. It enables the extraction of the displacement from the visual information. We used multivariate linear regression to obtain the PCM data corresponding to the displacement value. Displacement values and time were obtained by vibrating an object with the sound that is a pure tone, and PCM data of the sound were extracted to generate training data. Through the regression model trained in this method, PCM data corresponding to the displacement value was obtained to restore the sound. To evaluate the performance of our method, we measured the accuracy of restoration using energy spectrum density(ESD). Our experiment achieved a high accuracy of restoration of at least 85% on pure tones. And the accuracy of restoration at 72% and 75% of the talking sound and classic music "The carnival of animals".

KEYWORDS

Vibration, PCM data, Multivariate linear regression, ESD

1 INTRODUCTION

Humans hear sounds through the process of collecting auditory information from the eardrum, one of the sensory organs. Likewise, the eye receives light to collect visual information to see an object.[1, 2] However, when the eardrum or eye is damaged, the information collection function that the sensory organ is responsible for is deteriorated, causing great inconvenience in life. To reduce these inconveniences, the function of the sensory organs is compensated by wearing hearing aids for those with impaired eardrums and glasses for those with impaired eyes (cornea). However, if there is a defect in the sensory organs or it does not work, another method is needed.

In this work, we propose the method for converting initial information collected by sensory organs into different types of sensory information. The process of listening to sound, the initial information is heard using visual information rather than auditory information in this method. This allows the sound to be heard using only visual information, whether in an environment where the sound is not heard or in a situation where the eardrum is damaged. We call this method "Synesthetic information collecting method". This method can collect the desired information using different types of information, even if there is no initial information.

Considering the characteristics of sound waves transmitted through the vibration of the medium, the image was selected as the input of the synesthetic information collection method. Sound waves transmitted by the vibration of air vibrate the object when they touch them. This vibration contains the information of original sound waves. Using this characteristic, input sources were obtained by taking videos of the object vibrating. Then, videos were processed by phase-based video magnification to be efficient in other processes. The processed visual information was divided into frames and the displacement values of the vibrating object were extracted. This displacement data has information about the waveform of the original sound, but by itself does not match the waveform of the original sound. Therefore, regression

* Corresponding author

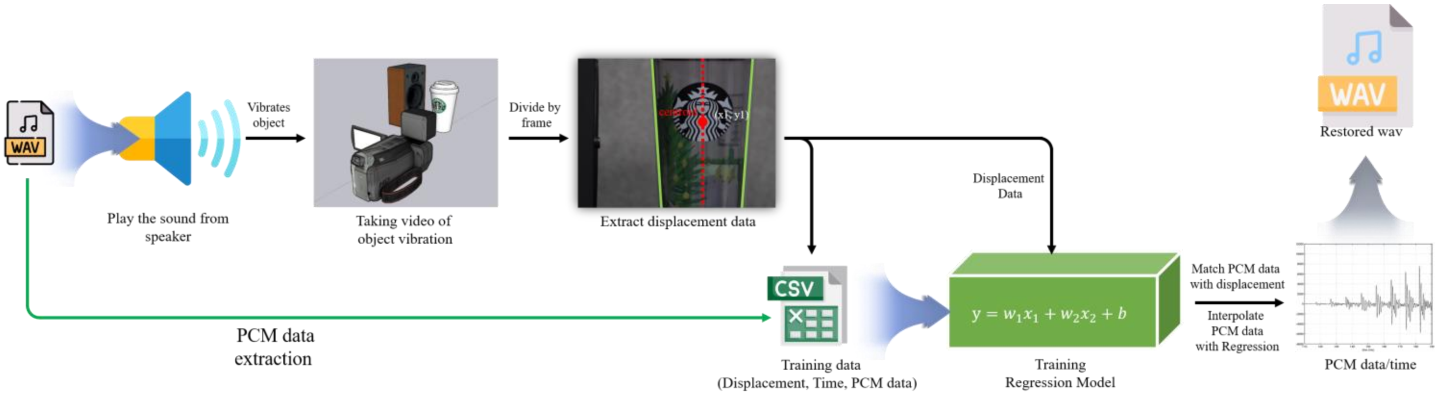
WAV
Play the sound from speaker
Vibrates object
Taking video of object vibration
Divide by frame
centroid
(x1, y1)
Extract displacement data
Displacement Data
WAV
Restored wav
PCM data extraction
CSV
Training data (Displacement, Time, PCM data)
$y = w_1x_1 + w_2x_2 + b$
Training Regression Model
Match PCM data with displacement
Interpolate PCM data with Regression
PCM data/time

Figure 1. Experiment process to restore the waveform of original sound

was selected as a method of generating PCM data, which is the displacement value of the speaker diaphragm according to time using displacement.[3, 10, 11] Using this regression model, it was possible to obtain PCM data of the restored sound corresponding to the distance. By arranging this PCM data in time, we could obtain the wav file of the restored sound. Then, the sensory conversion from sight to hearing was performed.

The new information collection method proposed in this paper is through visualization of auditory information. We suggest a new paradigm that is free from the limitations of sensory organs. Through this, we propose a method to collect information without inconvenience to people who have defects in sensory organs such as eyes or eardrums. And it suggests the possibility of more efficient and convenient information collecting without limiting the sense organs.

2 Method

2.1 Visualized auditory information and analysis

The purpose of this paper is to create a restored sound with information of the original sound by generating a waveform from the displacement of an object that has been vibrated due to the vibration of sound. It can be achieved through the visualization of auditory information. When the original sound, which is auditory information is played on a speaker, the vibration of air causes the surrounding objects to vibrate. At this time, since the vibration of the object is generated by sound energy transmitted by the original sound, it contains the information of the original sound. Based on this fact, this paper intends to restore the waveform of the original sound by processing or analyzing this vibration.

The vibration generated in this way is recorded as a video and auditory information is visualized. However, this information is so minute that neither the eye nor the computer can perceive it. Therefore, we want to use phase-based video magnification.[4] Phase-based video magnification is a technology that detects a phase optical flow through an image pyramid and expands the motion by enlarging the corresponding phase.[5, 6] Since the imaging in this experiment took micro-vibrations of an object at a fixed camera position, it would be more effective in the validity of the magnified result than using Eulerian video magnification that detects and enlarges changes in brightness.[7] Through this, the minute vibration caused by sound waves is enlarged, and visual information is processed for easier analysis and waveform restoration process.

2.2 Extracting displacement

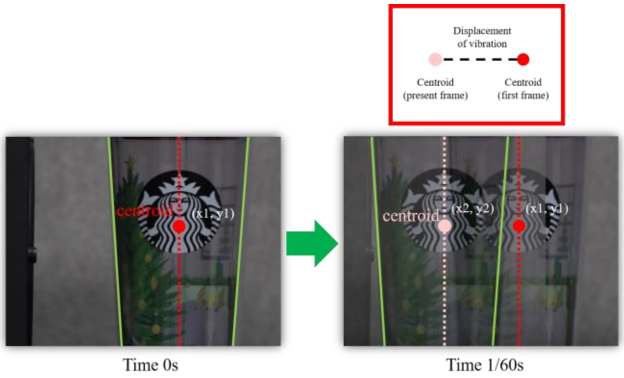

Figure 2. The method of extracting displacement

Through video magnification, it is possible to extract displacement from the vibration of the magnified object. As shown in Figure 1, the corresponding displacement value extracts the contour using segmentation of the vibrating object and background and extracts the centroid of the object. The first frame is set to the idle state, that is, the displacement is 0, and the measured distance between the centroid coordinates of the contour of the current frame is set as the displacement. The sign of the displacement value was set according to the phase of the original sound.

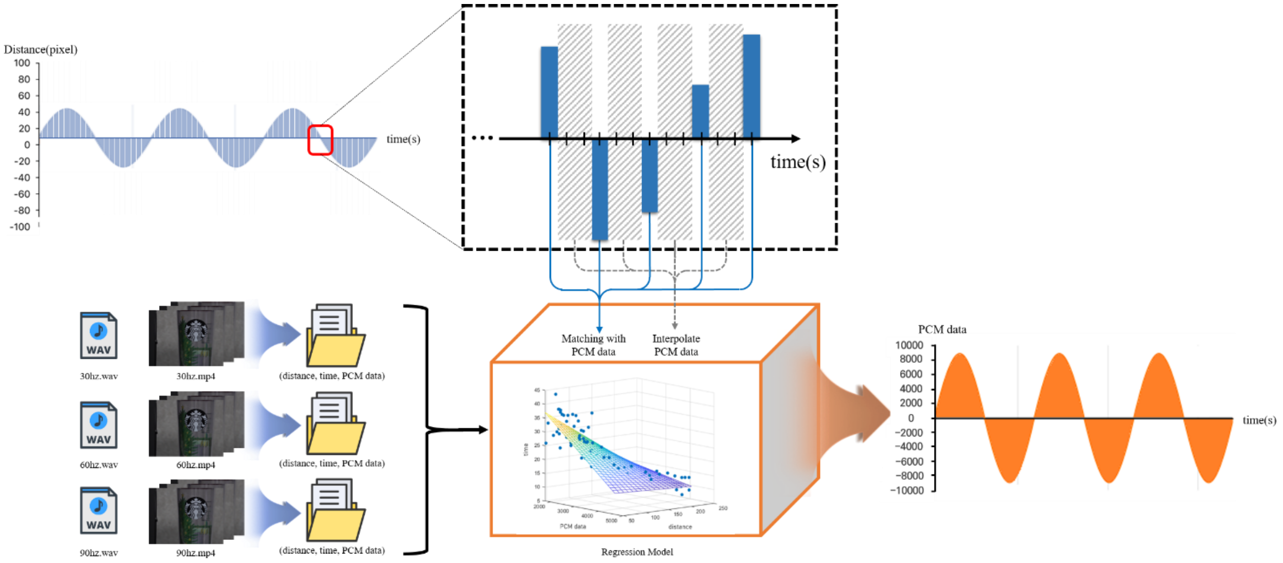

Figure 3. Process of restoring waveform by using regression model & Training regression model

2.3 Displacement - PCM data Regression model

The displacement value for each frame obtained by the process above cannot be called the PCM data of the original sound, that is, the waveform when it is arranged according to time. It is because of 1) the amplitude is not correct, and 2) the occurring difference in the fps value, which is the frequency at which the camera visualizes auditory information and samples (take a video) compared to the sample rate of the original sound. Therefore, we design a regression model that derives PCM data corresponding to the displacement and time values as factors as shown in Figure 3. For the corresponding regression model, the displacement value of vibration is obtained by playing sounds of several frequency bands, time and PCM data are extracted and used as training data.

Based on the regression expression learned and derived through this method, the corresponding PCM data can be got using the displacement and time values. This makes it possible to make a reasonable inference between the vibration of the object and waveform, thereby making the restoration process more accurate.

The value of unsampled frames is predicted with a distance value having a range value between the previous frame and the subsequent frame, and the effect of interpolating the PCM data value can be obtained by regressing with the time value. Based on the PCM data for the time thus obtained, a waveform of the restored sound and a wav file of the restored sound is generated, thereby completing the auditory visualization process.

3 Experiment & Result

3.1 Visualized auditory information processing and analysis

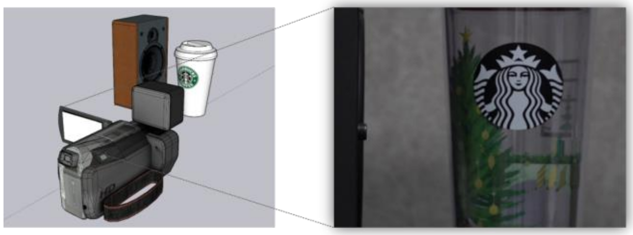

Figure 4. 3D line up of Experiment(Left) and view of camera(Right)

In order to achieve the sub-purpose of "visualization of auditory information", we designed the experiment as Figure 4. A tumbler, which is commonly seen in everyday life, was selected as an object to vibrate by sound waves. In the experiment, pure tone with a sample rate of 11025Hz and constant frequencies of 30 Hz, 45 Hz, 60 Hz, 90 Hz, and 120 Hz were generated for a total of 180 seconds. In the experiment of restoring the sound that is pure tone, it was possible to prove the validity of the research method proposed in this paper. In addition, the original sound that was downscaled to 11025 sample rate of the Talking sound and music "The Carnival of the Animals" was played for 180 seconds. These sound are used to check whether the information collection method works effectively even for original sound with complex waveforms.

The experiment proceeded as follows. It consists of a speaker to play the original sound, a tumbler (the object to vibrate) located close to the output of the speaker, and an EOS R5 camera to take it. In addition, the experiment was conducted in a soundproof environment and the intervention of external factors such as wind and noise was minimized by locating the speaker to play the original sound and the object to vibrate close. The video was taken on FHD(1920x1080 pixels) resolution with 120fps.

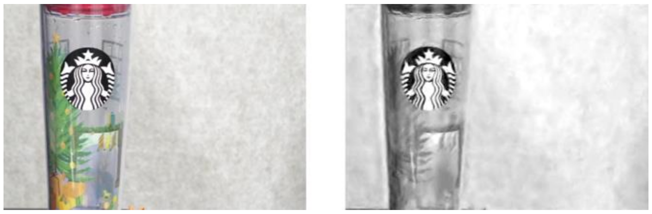

Figure 5. Capture of video that is original(Left) and magnified video(Right)

Figure 5 shows that phase-based video magnification is applied to the captured image to detect the phase difference caused by minute vibration of an object by sound waves and enlarge it to a range that can be checked with the naked eye. By doing this, the initial information was pre-processed into visual information that is efficient for displacement extraction and analysis processes.

3.2 Extracting displacement

Displacement should be extracted from the pre-processed visual information through process 3.1. For this, the contour between the object and the background was detected, and the vibrating object was detected.[8] Since the first PCM data of the origianl sound is 0, the centroid of the detected object in the first frame is set to idle. The euclidean distance between the centroid of the detected object of first frame and in the subsequent frame is measured, and this measured is displacement.[9] The displacement values of all frames were measured through the following equation.

$$\textbf{Displacement} = \sqrt{(x_n - x_0)^2 + (y_n - y_0)^2} \quad (1)$$

$$centroid\ at\ time\ n\ sec = (x_n,\ y_n)$$

$$centroid\ at\ time\ 0\ sec = (x_0,\ y_0)$$

Since each original sound was 30 seconds long, 3600 displacement data could be obtained.

3.3 Displacement - PCM data Regression model

In order to obtain PCM data using the measured displacement, a multivariate linear regression model that derives PCM data through displacement and time was constructed.[10, 11] The displacement values obtained through process 3.2 are 3600 for each sound source, and since a total of 5 sound sources (30Hz, 45Hz, 60Hz, 90Hz, 120Hz) of constant frequency were tested, a total of 18000 pieces of (displacement, time) data exist. By extracting PCM data of the sound source corresponding to each data, 18000 training data in the form of [(displacement: time), PCM data] were generated. In this way, a multivariate linear regression model optimized for the vibration of an object was trained to create a model that generates PCM data corresponding to displacement. The formula of the regression model is as follows.

$$y = w_1x_1 + w_2x_2 + b \quad (2)$$

$$y: PCM\ data$$

$$x_1: displacement$$

$$x_2: time$$

Then, to verify the validity of the trained model, a test set was generated using displacement: time data corresponding to unsampled PCM data. When the difference between the PCM data derived from the data of the test set and the original sound is less than 5%, The correct regression and its accuracy was measured. The results are shown in Table 1. That the higher the frequency of the test set, the lower the accuracy is. Such tendency is a problem itself that needs to be solved because with higher frequencies, more displacement values cannot be sampled when recording a video, leading to insufficient learning. For all that, in general, they all show high accuracy. Therefore, this regression model enables a reasonable derivation process between displacement and PCM data.

Table 1. Results of Accuracy test

Test Sound	Regression Accuracy
30Hz	95.56%
45Hz	90.10%
60Hz	91.26%
90Hz	89.52%
120Hz	92.28%

Using the regression model learned in this method, a restored sound was generated. Although there is displacement data obtained through processes 3.1 and 3.2, it is not sufficient for the sample rate of the original sound(11025Hz), so the values are filled linearly by considering the gradient between the previous value and the subsequent value. Through this method, an interpolating process was performed to match the number of PCM data and displacement data of the original sound. The formula for this process is as follows.[12]

$$x_k = gradient_{n:(n-1)}\ * k \quad (3)$$

$$(\mathbf{n-1 < k < n})$$

Then, PCM data is extracted with the learned regression model and listed according to time. The total number of PCM data was 11025Hz * 30s = 330750, and the wav file was created by composing it corresponding to the time. Through this, the information collection method through auditory visualization was completed. Figure 7. is the spectrogram of a 120Hz original sound and its restored sound.

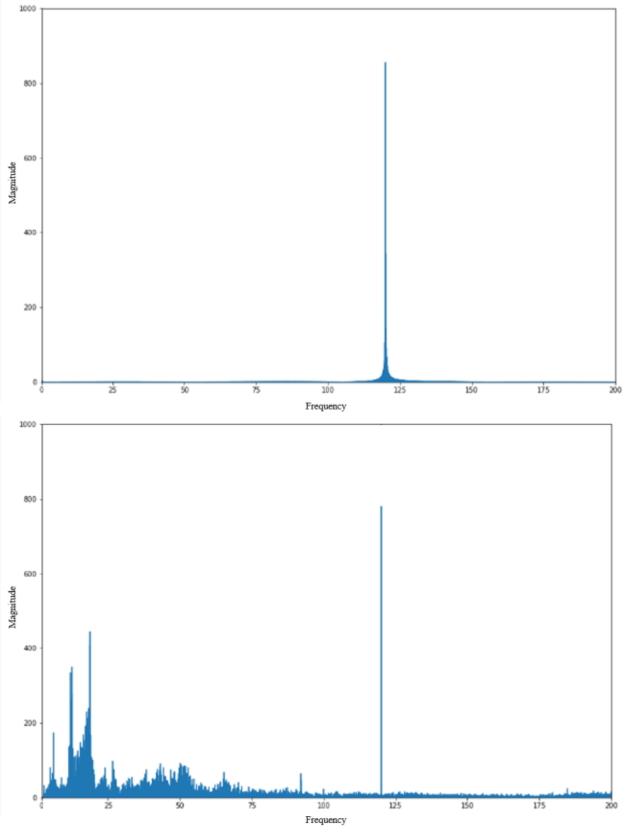

Figure 7. The spectrogram of the original 120Hz sound(top), and The spectrogram of the restored 120Hz sound(bottom)

3.4 Evaluating Accuracy of Restoration

To analyze whether the restoration process is reasonable, it is necessary to compare the restored sound with the original sound. For this, we use Fast Fourier Transform(FFT). The FFT transforms information in the time domain(sound waves) into the frequency domain called periodic functions.[13, 14] Energy Spectral Density (ESD), which is unique information in the frequency domain, is used as an index to compare the original sound with the restored sound.[15, 16] The formula to calculate ESD is as follows.

$$\boldsymbol{ESD = \int_{-\infty}^{\infty} |X(f)|^2 df = 2 * \int_{0}^{\infty} x^2(t) dt} \qquad (4)$$

$\boldsymbol{X(f)\ is\ Fourier\ Transform\ of\ aperiodic\ signal\ x(t)}$

Since ESD is an index indicating how much energy is located in a corresponding frequency band, it is possible to calculate the accuracy of restoration by comparing the ESD for all frequency bands of the original sound with restored sound. Therefore, the accuracy of restoration was measured in the following way.[17]

$$\boldsymbol{Restoration\ accuracy} = \frac{\boldsymbol{ESD\ of\ restored\ sound}}{\boldsymbol{ESD\ of\ original\ sound}} \times \boldsymbol{100}\ (\%) \qquad (5)$$

Table 2. The accuracy of restoration corresponding to original sound

Original Sound	Accuracy of Restoration
30Hz	93.72 %
45Hz	94.33 %
60Hz	85.67 %
90Hz	91.12 %
120Hz	90.02 %
Talking sound	72.27 %
The Carnival of Animals	75.52 %

Table 2 shows the results of accuracy of restoration. In the experiments of original sounds with a constant frequency, the accuracy of restoration was also high because the accuracy of the regression model was high overall. However, as the frequency increases, the accuracy of restoration tends to decrease. This is analyzed because as the frequency increases, the number of inflection points increases, and the effectiveness of the interpolation method decreases. However, when looking at the waveform of the restored sound, its frequency also has a value very similar to that of the original sound. Through this, it was possible to obtain the result that the experimental method of this paper is valid.

In the experiments of Talking sound and classical music "The Carnival of the Animals", the accuracy of restoration was low. As a result of actual listening, original sound and restored sound were so different. In the case of a music and talking, Since there are unique characteristics of pronunciation or an instrument, the restored sound is not perceived well when the shape of the waveform is not restored almost completely. However, this is expected to be improved by generating training data with original sound of more diverse frequency bands and creating more improved regression models.

4 CONCLUSIONS

In this paper, we propose a method of converting the initial information collected by sensory organs into other types of sensory information. Considering the characteristics of sound waves that are transmitted through the vibration of the medium, a video was selected as the input of the synesthetic information collection method. Sound waves transmitted by the vibration of air, which is a medium, vibrate when it hits a nearby object. At that time, the vibration contains the information of the initial sound wave as it is. Using these features, the input sources were obtained by taking the vibration of an object as a video. The initial visual information was pre-processed by phase-based video magnification. Then, the displacement value was extracted to generate a restored sound. The multivariate linear regression model was created with training data created using this displacement value and PCM data of the original sound. Using this regression model, the displacement could correspond to PCM data. In addition to this, the interpolation of the displacement compensated for the insufficient number of data. it allowed an effective and valid derivation of displacement to PCM data. Finally, the restored sound could be generated. By comparing the ESD of the restored sound and the original sound, we were able to measure the accuracy of restoration. high accuracy was obtained in the experiment of the original sounds that have a constant frequency. Music and Talking sound with complex waveforms showed low accuracy of restoration, and it was difficult to find similar parts in actual listening.

The results of this paper are significant in that we proposed a new paradigm of collecting sensory information. Previously, it was impossible to hear sound unless the initial information was sound waves. However, the experiments in this paper show that auditory information can be collected even when initial information is visual information. Therefore, the results of this paper will be helpful in resolving the dysfunction or loss of sensory organs responsible for hearing and vision. And it will be an essential key to removing the limitations in the collection of sensory information.

ACKNOWLEDGE

This work was partly supported by the Technology development Program of MSS (S2962806) and the IITP grant funded by the Korea government (MSIT) (No.2019-0-01842, AI Graduate School Program (GIST)).

REFERENCES

[1] Artal, Pablo, and Antonio Guirao. "Contributions of the cornea and the lens to the aberrations of the human eye." Optics Letters 23.21 (1998): 1713-1715.

[2] Lin, Gau-Feng, and Joel M. Garrelick. "Sound transmission through periodically framed parallel plates." The journal of the acoustical society of America 61.4 (1977): 1014-1018.

[3] Serra, Xavier, and Julius Smith. "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition." Computer Music Journal 14.4 (1990): 12-24.

[4] Wadhwa, Neal, et al. "Phase-based video motion processing." ACM Transactions on Graphics (TOG) 32.4 (2013): 1-10.

[5] Gautama, Temujin, and M. A. Van Hulle. "A phase-based approach to the estimation of the optical flow field using spatial filtering." IEEE transactions on neural networks 13.5 (2002): 1127-1136.

[6] Adelson, Edward H., et al. "Pyramid methods in image processing." RCA engineer 29.6 (1984): 33-41.

[7] Gautama, Temujin, and M. A. Van Hulle. "A phase-based approach to the estimation of the optical flow field using spatial filtering." IEEE transactions on neural networks 13.5 (2002): 1127-1136.

[8] R. -W. Bello, A. S. A. Mohamed and A. Z. Talib, "Contour Extraction of Individual Cattle From an Image Using Enhanced Mask R-CNN Instance Segmentation Method," in IEEE Access, vol. 9, pp. 56984-57000, 2021, doi: 10.1109/ACCESS.2021.3072636.

[9] Wang, Liwei, Yan Zhang, and Jufu Feng. "On the Euclidean distance of images." IEEE transactions on pattern analysis and machine intelligence 27.8 (2005): 1334-1339.

[10] Leggetter, C. J., and Philip C. Woodland. "Speaker adaptation of continuous density HMMs using multivariate linear regression." ICSLP. Vol. 94. 1994.

[11] Weisberg, Sanford. Applied linear regression. Vol. 528. John Wiley & Sons, 2005.

[12] Narang, Sunil K., Akshay Gadde, and Antonio Ortega. "Signal processing techniques for interpolation in graph structured data." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[13] Bracewell, Ronald Newbold, and Ronald N. Bracewell. The Fourier transform and its applications. Vol. 31999. New York: McGraw-Hill, 1986.

[14] Nussbaumer, Henri J. "The fast Fourier transform." Fast Fourier Transform and Convolution Algorithms. Springer, Berlin, Heidelberg, 1981. 80-111.

[15] Hioka, Yusuke, et al. "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain." IEEE Transactions on Audio, Speech, and Language Processing 21.6 (2013): 1240-1250.

[16] Menounou, Penelope, and David T. Blackstock. "A new method to predict the evolution of the power spectral density for a finite-amplitude sound wave." The Journal of the Acoustical Society of America 115.2 (2004): 567-580.

[17] Saputra, Laurentius Kuncoro Probo, Hanung Adi Nugroho, and Meirista Wulandari. "Feature extraction and classification of heart sound based on autoregressive power spectral density (AR-PSD)." 2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering. IEEE, 2014.