Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776802

# AN AUDIO QUALITY IMPROVEMENT ALGORITHM FROM LOW RESOLUTION INPUT USING DEEP LEARNING TECHNOLOGY

**Hong-Jin Kim [1],Jeong Tak [2] Ryu, Kyuman Jeong [1]**
[1] School of AI, Daegu University, Gyeongsan-si, Korea.
[2]College of Information and Communication Engineering, Gyeongsan-si, Korea.

**Abstract:** Artificial intelligence technology, in which computers perform actions or behaviors similar to humans, is becoming popular. In particular, many efforts have been made to implement a technology that distinguishes objects or responds to user actions. Furthermore, it is also in the limelight in fields that require a lot of time and effort, such as restoring paintings drawn in the past. It is expected that it can be used in various fields as well as image restoration techniques using 3D data. In particular, audio data has changed from the method of using physical storage devices in the past to the form of being provided on a network basis, and the future market value is also great. In this paper, we propose an algorithm to restore compressed audio data so that it can self-produce high-quality audio data from an internal storage device. We propose a method of restoring audio data that is reproduced by enumerating one-dimensional data that changes over time using lossless audio data and audio data lost after compression through a convolutional neural network (CNN), a deep learning technology.

**Keywords:** Artificial Intelligence, CNN, Audio Super-Resolution, Deep Learning

## 1. Introduction

Recording information makes the greatest contribution to the development of human technology. In the past, it was common to pass down texts or pictures to future generations, but in modern times, as sound and images can be recorded, various information can be easily accessed. In particular, the act of recording sound rather than video has evolved from gramophones to LPs, cassette tapes, CDs, and MP3s to be gradually miniaturized and easily transmitted. The biggest reason for being able to move from the material realm to the intangible is the recording method in a file format that can be easily transmitted. As the penetration rate of the Internet and smartphones rapidly increased, the way data was used using physical storage devices in the past has changed. In particular, it has become possible to receive more data through the Internet than a storage medium that a mechanical device can hold, such as a network-based cloud service, and streaming services that wirelessly provide video or sound source data are in the limelight.

[1] Corresponding author: kyuman.jeong@daegu.ac.kr

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A07091326)

In this research work, in order to solve the size and time constraints in Internet-based data transmission, we will reduce the size of transmitted data and present a technology that enables efficient audio streaming through self-restoration in the mechanical device responsible for physical storage and execution.

## 1 2. Related Research

In order to realize Audio Super-Resolution, it is necessary to understand how to process existing standardized audio data. Audio compression methods include a lossless compression method with matching characteristics on the waveform using signal waveform characteristics and a human auditory. It can be divided into a lossy compression technique that compresses an audio signal up to a limit so that there is no difference. In order to restore lossy compressed data, which is the goal of this paper, it is necessary to understand the concept of a standardized lossy compression method.

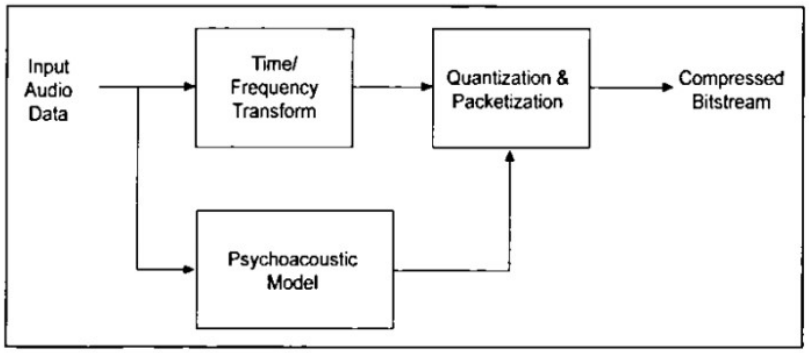



Figure 1: Conceptual diagram of audio compression method using psychoacoustic model (excerpt from [1])

Among technologies for restoring data, there is a method using machine learning. Machine learning is a learning algorithm based on biological neural networks. When using machine learning, artificial neurons are used as the minimum unit. A network structure in which a plurality of neurons are used and combined is called an Artificial Neural Network (ANN). ANN has a structure that repeats Linear Fitting and Nonlinear Transformation when classifying data with the motif of human neural structure.

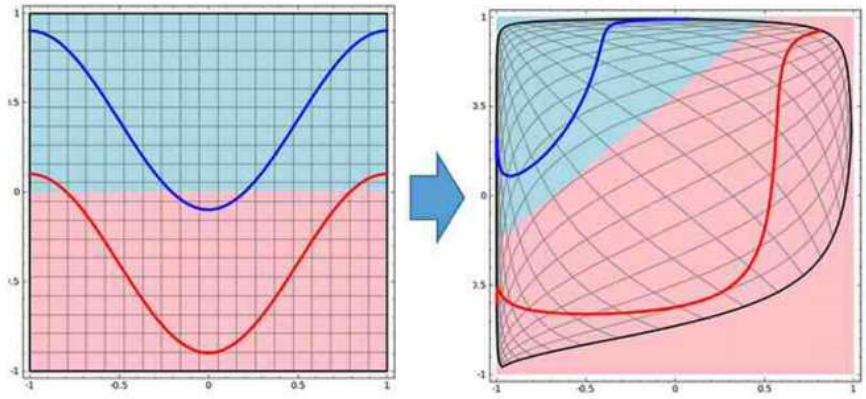


Figure 2: Linear fitting and nonlinear transformation of artificial neural networks (excerpt from [2])

Data classification is the most representative method using artificial neural networks. As shown in Figure 2, when designing an algorithm that obtains a line that can best distinguish the data of the blue and red areas, errors may occur in some areas converging to 0 in the graph, as shown on the left, but distortion as shown on the right. If the spatial graph is used, more sophisticated data classification is possible. Perceptron is composed of input and output layers and uses a method of learning the weight between neurons in each layer.

Convolutional Neural Network (CNN) is a field of deep learning that uses supervised learning. CNN, a model capable of recognizing patterns even when patterns are changed by imitating the visual processing process, was born in 1998 as a proposal by Yann Lecun [14], and has been recognized in image-related fields such as winning the image recognition contest called ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

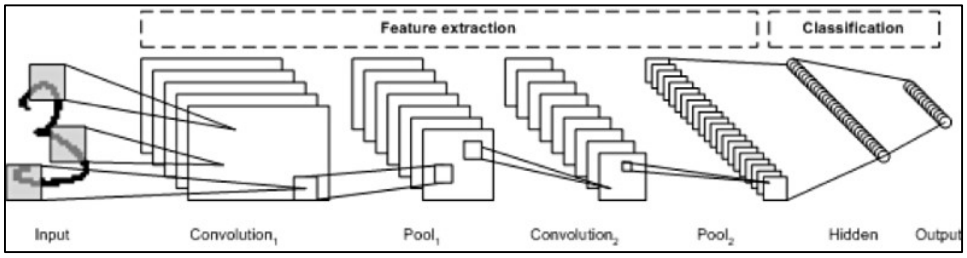



Figure 3: Convolutional Neural Network

Figure 3 explains the CNN structure. In feature extraction to extract features from the input data, convolution and pooling steps are repeated, and when the final pooling layer is reached, classification is performed to derive the result.

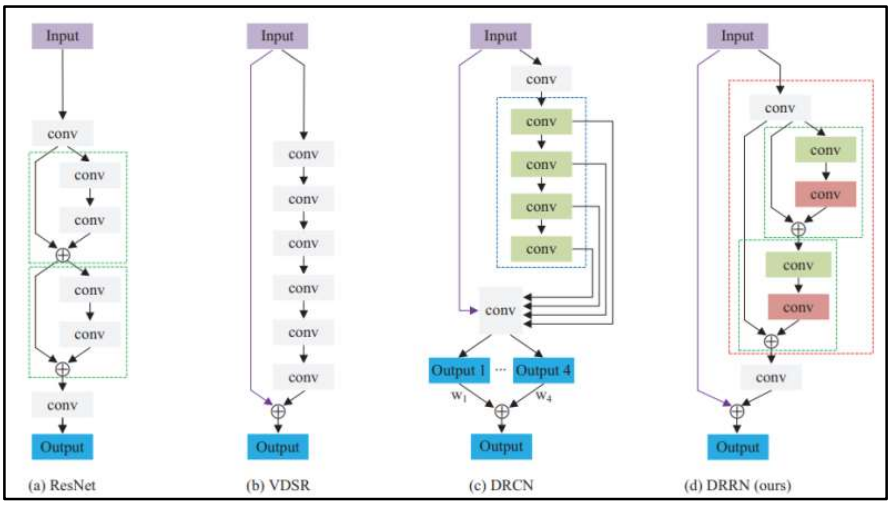



Figure 4: Layer structure of (a) ResNet, (b) VDSR, (c) DRCN, and (d) DRRN (excerpt from [9])

CNN models used for Super-Resolution include ResNet, VDSR [10], DRCN [13], and DRRN [9]. ResNet is a model developed by Microsoft researchers. In a general deep learning structure, it is difficult to expect effective learning when the layer structure deepens. It was solved with Gradient Clipping [12], which is one of the methods to prevent the divergence of the gradient, which indicates the update direction of the parameter. VDSR uses an algorithm

that constructs a very deep layer with a pair of Convolution-Layer and ReLU activation function and adds the result and input of the last layer element by element. DRCN solves the overfitting problem by recursively learning in one layer, reduces the number of parameters that increase together as the learning layer increases, and makes it possible to learn up to 20 layers. DRRN uses an algorithm that learns by layering recursive blocks and adds each result with the initial layer.

## 2 3. Proposed Method

In this research paper, we study how to convert low-quality audio data into high-quality data using CNN used for media restoration. In the process of restoring sound quality using deep learning, WAV format data extracted from CD is converted into low-quality data and high-quality data, and used as input values and GT (GroundTruth) to be used for learning, respectively. We want to design an algorithm that restores audio data through repeated learning. The basis of the model used for learning is to compare and evaluate various models using CNN to select the model with the highest degree of completeness, and to introduce the process of designing a new model by modulating the model if necessary.

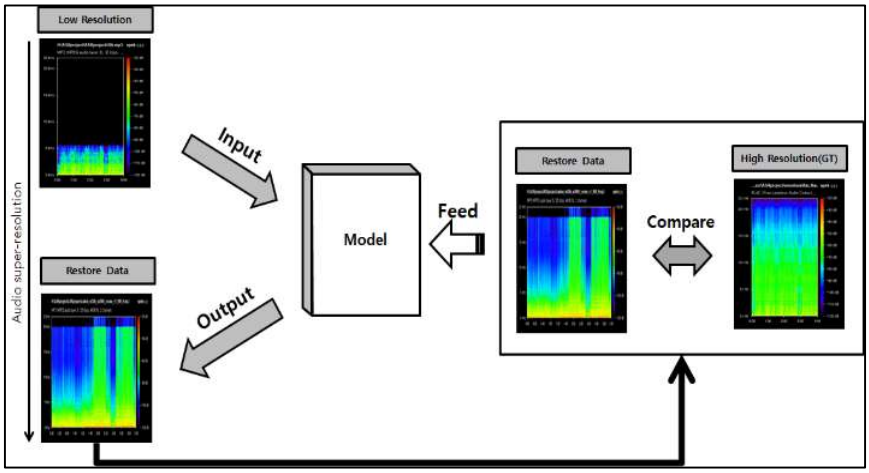



Figure 5: Audio Quality Restoration Algorithm

### Data set generation

In the learning data generation process, 101 sound source CDs are used, and since each audio file has a different playback time, it is necessary to establish a unified standard. The data generation process is a total of two steps, extracts the audio data stored in the first CD, converts it into a lossless compression format Flac, which is the standard of the restoration point, and saves it. Next, in order to generate low-bitrate audio data to be used as input data, the Flac file extracted from the CD is down-sampled and generated. The number of extracted sound sources is a total of 1,137, and the compressed data must be decoded using the Pydub library that supports FFmpeg (Febrice Bellard et).

A pre-processing process is required to learn the training data into the model. When converting 44,100 Hz sample rate sound source data into 1-channel mono, 16-bit raw data, assuming that each sound source is 3 minutes, an array of 7,938,000 size can be extracted.

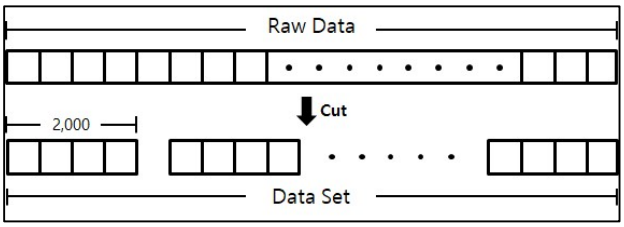



Figure 6. Split dataset for training

When 1137 sound sources to be used in the experiment are stored in an array, it can be converted into about 9 billion arrays, and the extracted array is divided into 2000 units, which is a size suitable for deep learning, and stored.

In the case of a general sound source file, it is composed of 2-channel stereo, and left and right data are alternately arranged at 16-bit intervals, so they are stored separately.

## Learning model

In this research work, we implement and implement a method to restore audio quality through extension by controlling the audio bandwidth using deep learning. The methods used for implementation were DRRN (Deep Recursive Residual Network) and SRResNet (Super-Resolution Residual Network) models used in Image Super-Resolution. Unlike image data, audio data stored in a one-dimensional array is used. Therefore, a new model is designed and proposed to have higher long-term accuracy, and it undergoes a process of comparative analysis with various previously presented CNN models. Compared with ResNet and DRRN, our proposed model is shown in Figure 5, and RES_UNIT uses the Residual Unit structure of the DRRN algorithm. For the initial input, not a single residual unit, but multiple residual units are used to extract features or patterns of different data, and convolution is performed to the next layer by adding two equation pairs to the passed results, and the last one is passed through the residual unit. One result is passed through a single convolution map to derive the output.

The Audio Super-Resolution technique proposed in this paper and the existing Image Super-Resolution show differences in the structure of data used for learning. Audio data has a 1-dimensional 16-bit array structure unlike images that have 3-dimensional array information, so the data in the learning model is reduced from 3-dimensional to 1-dimensional. As the data is reduced from 3D to 1D, it has an advantage over images in terms of computational speed, but since the size of the filter is reduced, it is necessary to modify the model to correlate with the surrounding data.

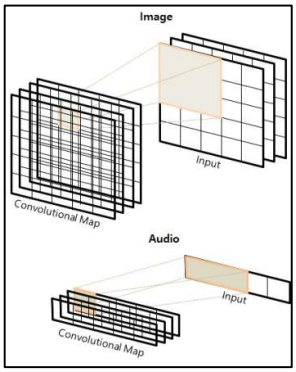
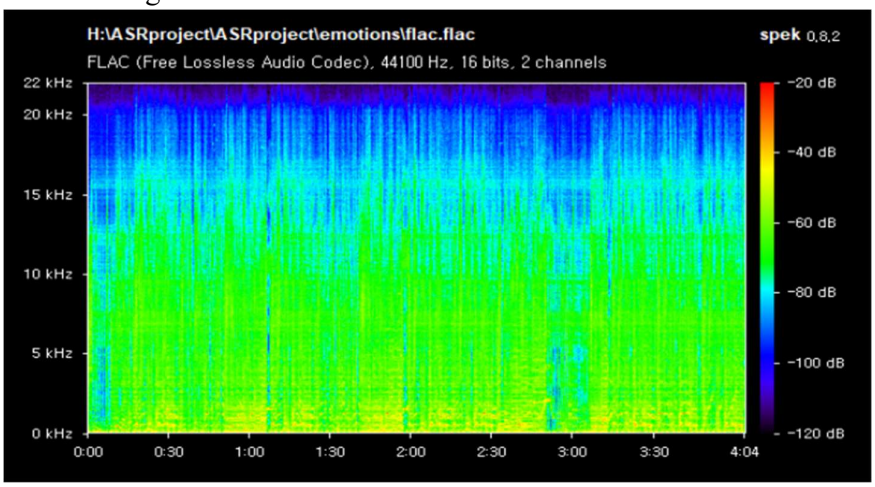


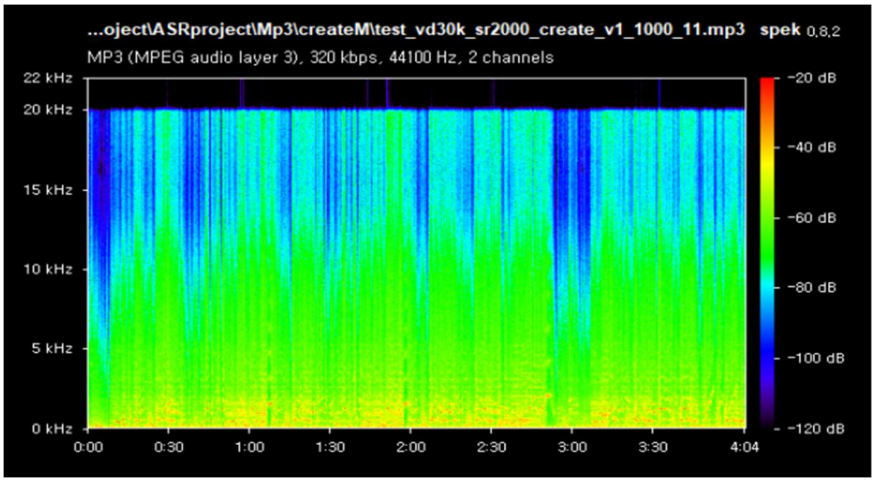
Figure 7: Convolutional difference

## 3 4. Experiments and Results

After dividing the total 1,137 raw data by channel, it was studied using 4,512,753 datasets of High Resolution data and Low Resolution data, respectively, prepared by dividing into 2,000 sizes. The spectrum of the sound source to be used as the test set after learning is shown in Figure 8 below.




(a) Original sound




(b) Learning result

Figure 8: Spectrum for the original sound source (a) and the result obtained through training (b)

The learning results using ResNet, DRRN, and our proposed model are shown in Figure 9, Figure 10, and Figure 11, respectively.

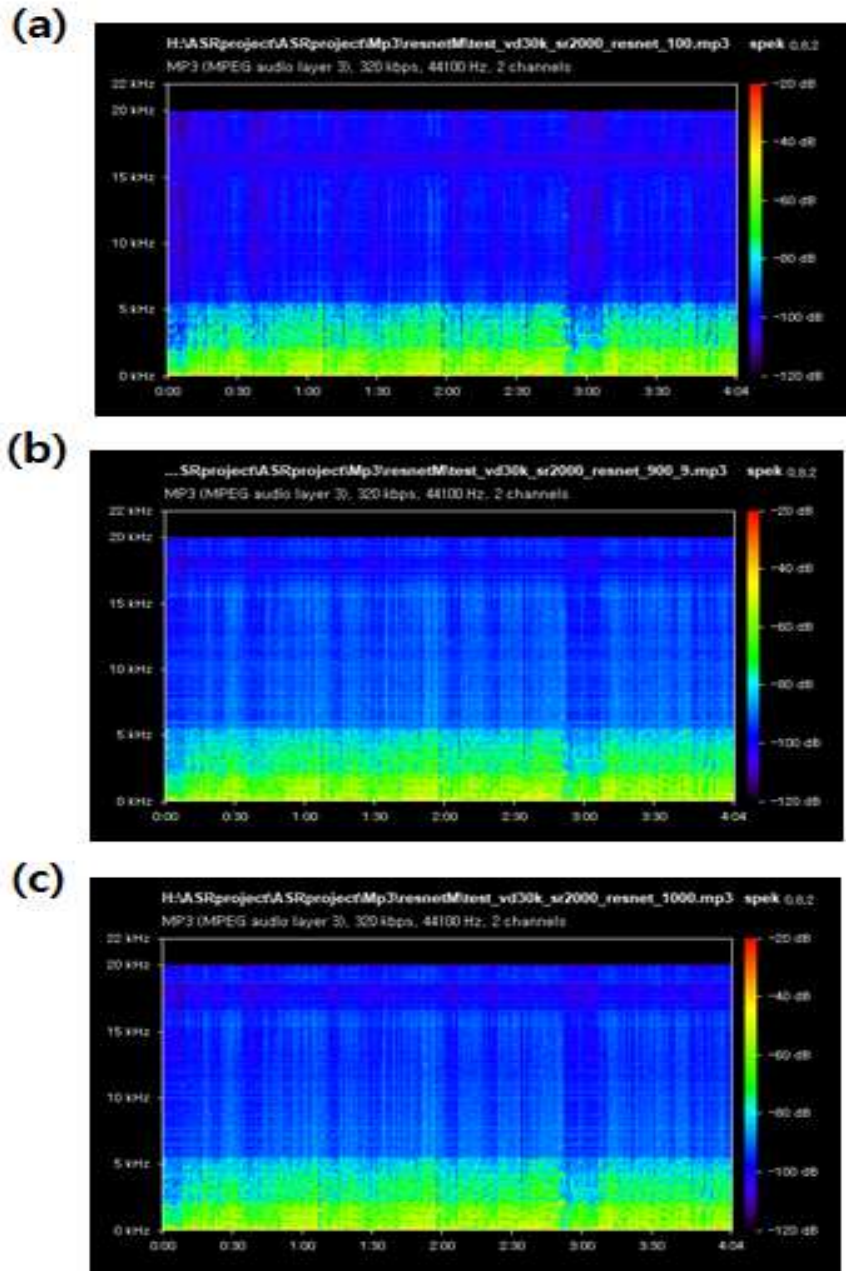



Figure 9: ResNet 100 (a), 900 (b), and 1000 (c) learning results

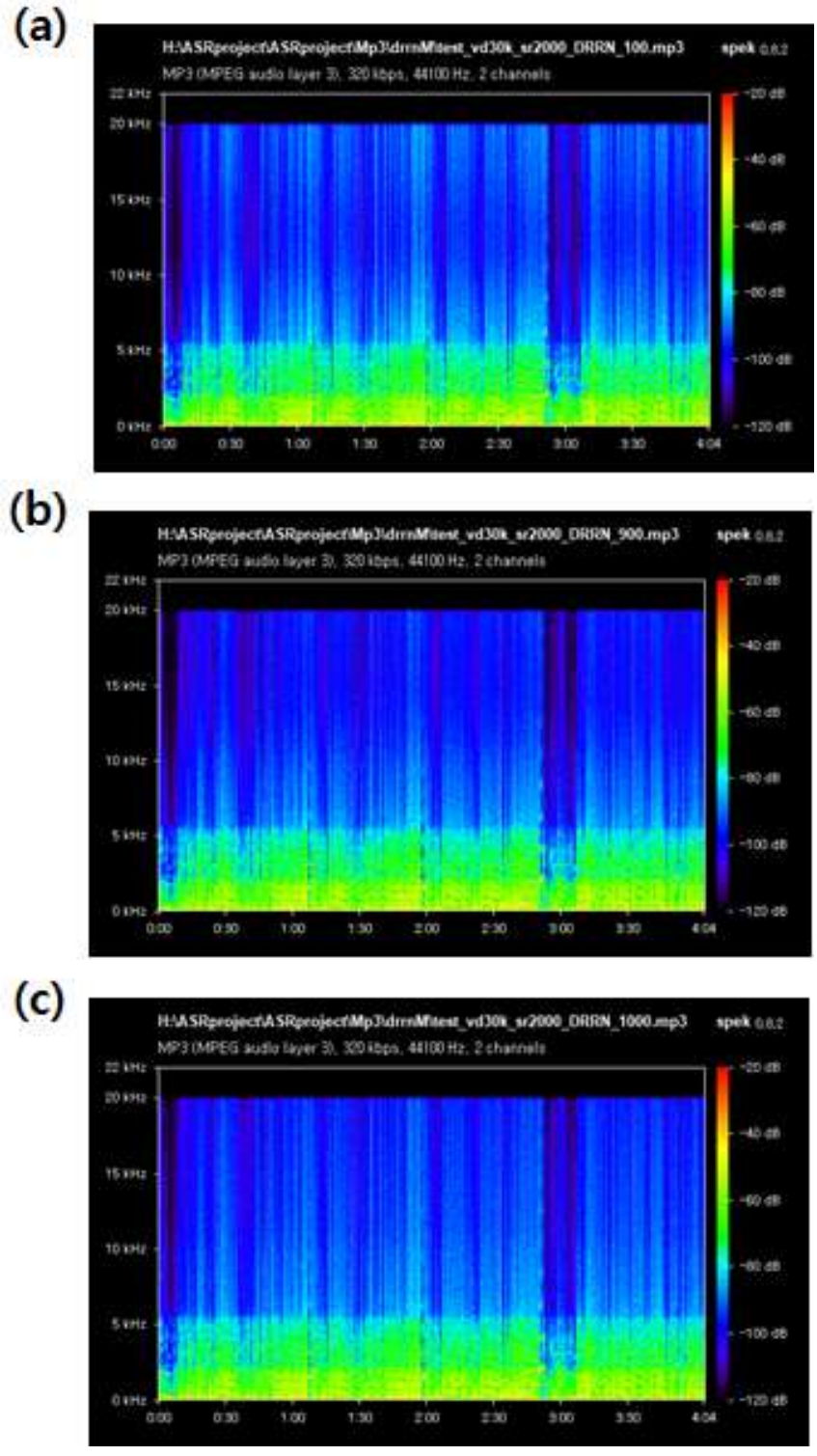



Figure 10: DRRN 100 (a), 900 (b), 1000 (c) learning results

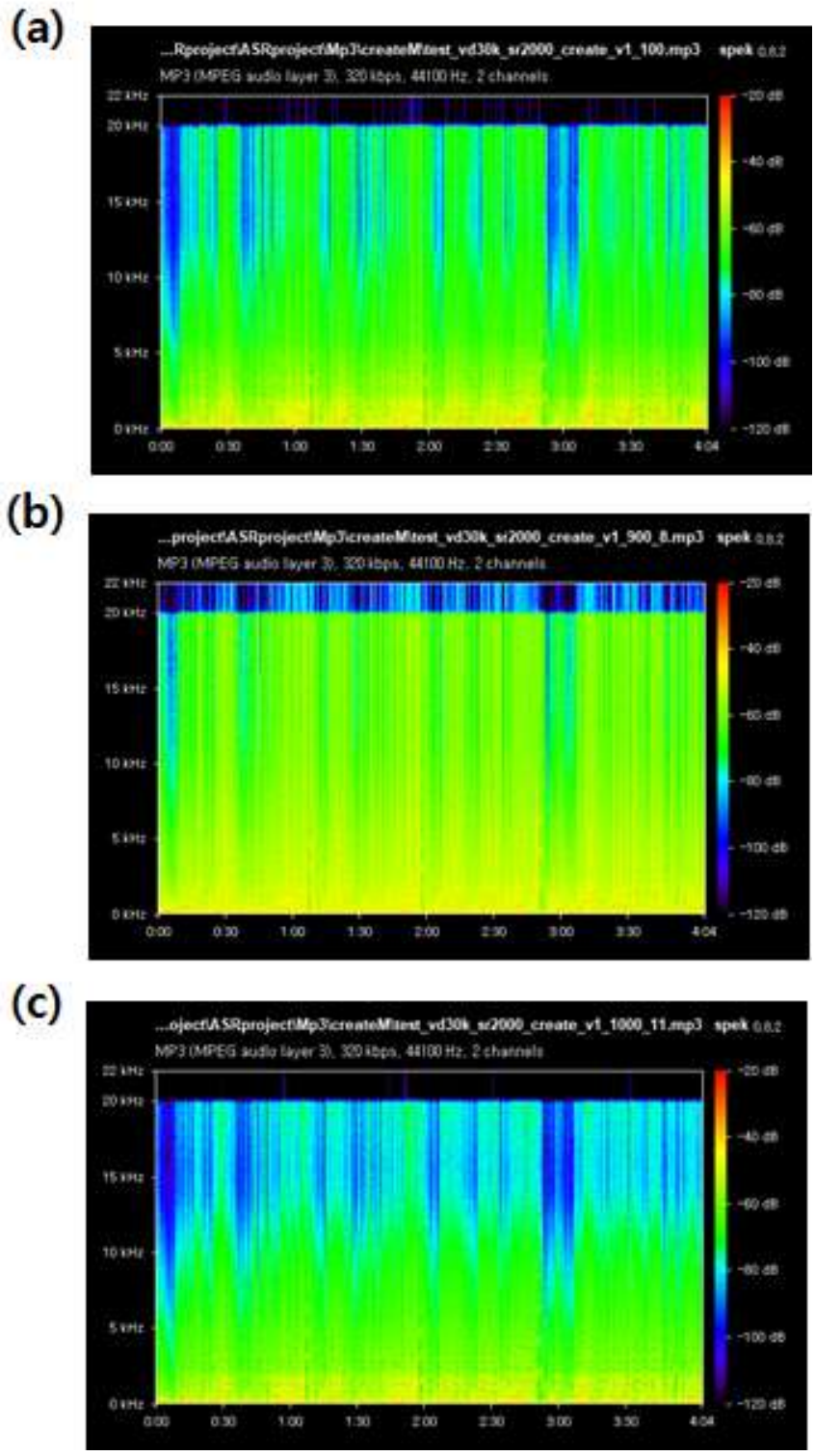



Figure 11: Results of learning the proposed model 100 times (a), 900 times (b), and 1000 times (c)

Table 1. Match ratio by model according to Train Step (closer to 0, better result)

| Model | Train Step | Loss |
|---|---|---|
| ResNet | 100 | 0.496398618 |
| | 900 | 0.502157432 |
| | 1000 | 0.500778312 |
| DRRN | 100 | 0.454160073 |
| | 900 | 0.470672504 |
| | 1000 | 0.502725412 |
| Proposed Model | 100 | 0.462948464 |
| | 900 | 0.345112004 |
| | 1000 | 0.227719983 |

Table 1 compares each result with the raw data of the original sound source and averages the differences. In the case of ResNet or DRRN, there is no significant difference, but in the case of our proposed model, it can be confirmed that good results are obtained with values close to 0.

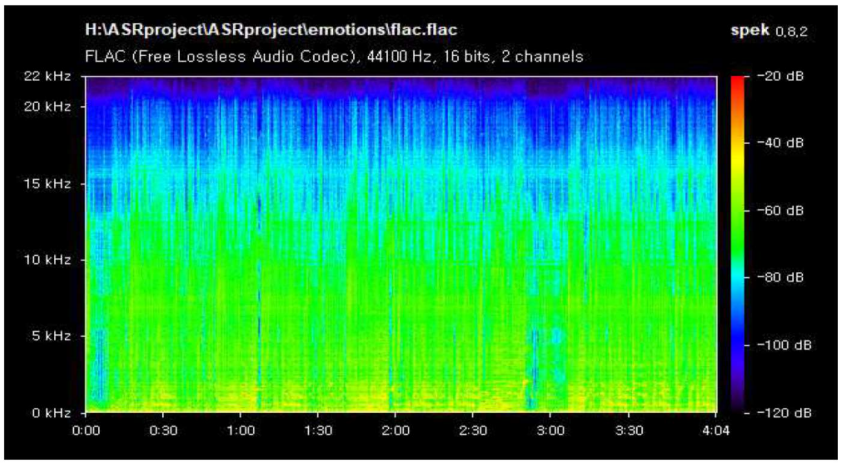
## 5. Conclusions

As models to be used in the training process, ResNet and DRRN, which are CNN models that are often used to restore high-resolution images from low-resolution, were used. When restoring the sound quality, it was confirmed that the model used in the existing Image Super-Resolution could not obtain good performance.

As shown in Figure 12, although the high-pitched data is alive, there is a problem that it does not match the original spectrum which is still different from the original one. In addition, noise generated during the restoration process can be identified. The time until the sound source derives the result through the model also takes about 3 to 5 minutes when using the current 3 minutes sound source.

Since audio has a lower data dimension than images, the proposed model uses a number of Residual Units at the initial input rather than the size or number of filters or the depth of the layer, and learns in pairs. It has been confirmed, and it is expected that better results can be obtained if the number of training times, the size and number of filters, and the size of the data to be learned are adjusted.

(a)




(b)

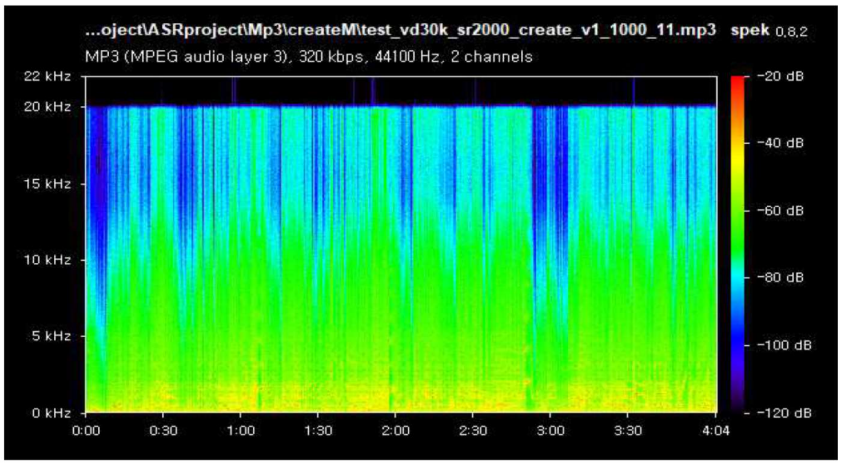



Figure 12: Spectrum of the original sound source (a) and the result obtained through learning (b)

## References

[1] Ying Tai, Jian Yang and Xiaoming Liu. (2017). Image super-resolution via deep recursive residual network. In CVPR

[2] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. (2016). Accurate image super-resolution using very deep convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition3.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

[4] V. H. Quintana and Edward J. Davison. (1974). Clipping-off gradient algorithms to compute optimal controls with constrained magnitude. International Journal of Control, vol. 20, no. 2, pp. 243-255.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems.

[6] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. (2016). Deeply-recursive convolutional network for image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[7] Joan Bruna, Pablo Sprechmann and Yann LeCun. (2016). Super-resolution with deep convolutional sufficient statistics. In International Conference on Learning Representations (ICLR).

[8] Volodymyr Kuleshov, S. Zayd Enam and Stefano Ermon. (2017). Audio super-resolution using neural nets. Presented at the 5th International Conference on Learning Representations (ICLR)

[9] Ying Tai, Jian Yang and Xiaoming Liu. (2017). Image super-resolution via deep recursive residual network. In CVPR

[10] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. (2016). Accurate image super-resolution using very deep convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition

[12] V. H. Quintana and Edward J. Davison. (1974). Clipping-off gradient algorithms to compute optimal controls with constrained magnitude. International Journal of Control, vol. 20, no. 2, pp. 243-255

[13] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee. (2016). Deeply-recursive convolutional network for image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[14] Joan Bruna, Pablo Sprechmann and Yann LeCun. (2016). Super-resolution with deep convolutional sufficient statistics. In International Conference on