1. Necessity of research projects

Speech Enhancement has a long history of research by removing noise from noise and clean speech so that listeners and machines can hear clean sounds. Traditionally, speech enhancement has been achieved by using techniques such as Filter Bank, or by attenuating a specific portion of the data using Discrete Wavelet Transform (DWT), which is impossible to learn about a given data. Such a method does not effectively eliminate new noise not considered at the time of development.

In order to solve these existing problems, learningable methods introduced deep learning have been introduced. Recently announced deep learning models vary depending on 1) domain differences in input data, 2) time-frequency domain conversion methods in time domains, 3) neural network structure differences, and 4) neural network loss functions.

Input Data Domain Differences: Traditionally, voice signal processing has been performed primarily in the time-frequency domain. However, recent research shows that it is difficult to say that either input data domain has an absolute advantage in the deep learning model domain, with time domain methodologies also recording state-of-the-art (SOTA) at the time of presentation by introducing appropriate neural network structures and loss functions.

Time-to-Frequency Domain Conversion Method: Time-to-Frequency Domain Conversion Method has traditionally been dominated by Short-Time Fourier Transform (STFT) and Mel Filter Cephal Coefficient (MFCC) methods. However, a method of introducing Continuous Wavelet Transform (CWT) is also being announced to solve the problem of predetermined frequency range according to the window size set during the conversion process and block development caused by discontinuity between window blocks.

Structure of speech-enhanced neural network models: Most of the structures of early speech-enhanced neural network models are based on encoder-decoder structures, but various model structures have been announced recently. For example, the encoder-decoder model has ResNet module and Attention module as a block instead of LSTM and stacks them up, has separate encoder-decoder paths for Magnitude and Phase in the time-frequency domain, has a connection path between the two encoder-decoder paths, and U-Net structures invented for medical image segmentation. The continuous presentation of new neural network structures means that a suitable structure for voice-enhanced neural network models has not yet been completed.

Loss function of speech-enhanced neural network model: Loss function is an important factor in training neural networks. However, it is not easy to directly define what neural networks want to learn. For example, when targeting the neural network in an encoder-decoder-based model as clean speech and applying the L1 loss function, achieving small losses does not guarantee better noise cancellation performance. Recently, Generative Adversarial Networks (GANs) has attracted attention for proposing a Metric loss function method that uses direct performance indicators as loss functions of neural networks. In the field of voice improvement, new attempts are also needed, such as defining evaluation figures directly related to voice improvement as direct loss functions.

Deep learning-based voice improvement research can be summarized into the following four, and these are topics that require continuous research.

· Analysis of Time Domain and Time-Frequency Domain Models

· A Study on the Time-Frequency Domain Data Acquisition Method

· A Study on the Structure of Neural Networks Suitable for Speech Enhancement

· A Study on the Loss Function Suitable for Speech Improvement

Meanwhile, the voice orientation may be used as a source technology for voice recognition, active magazines.They are a very high field of industrial utilization.In particular, voice recognition is expanding as a core technology such as smart speaker, and refrigerator, and refrigerator is drawing attention to user interaction.In addition, voice recognition technology may be used for user recognition.Active noise is a technology that reduces noise signal and reducing noise signals and the noise signal.The key of this technology is the same as the voice-oriented technology.This technology is used to reduce driving noise between soldiers in the electricity, electricity, driving noise reduction between soldiers in the car indoor, and cars.Voice orientation technology is a source technology as a source technology of voice recognition technology.

2. Objectives and contents of research projects

1) the ultimate goal

This research project aims to develop the highest performance deep learning-based voice improvement model. This objective is again subdivided into the development of a time domain model and a time-frequency domain model. First, the developing time domain model solves the problem of speech enhancement quality degradation caused by aliasing nested in downsampling and upsampling processes of encoder-decoder-based models. Second, the time-frequency domain model studies the application of continuous wavelet-based spectral decomposition methods to minimize information loss arising from the transformation process. Discrete wavelet-based spectral decomposition, which has been applied to speaker recognition, is not suitable for speech enhancement, which requires detailed handling of speech signals.

· Development of a Time Domain Voice-Enhancing Neural Network Model of Encoder-Decoder Structure Using Anti-Aliasing

· Development of a Time-Frequency Domain-Based Voice-Enhanced Neural Network Model Using Continuous Wavelet Transform (CWT)-Based Spectral Decomposition

2) Research objectives and research contents for the first year

[Research goal for the first year]

· We achieve SOTA performance in the time domain model.

· To this end, the model structure solves the Aliasing problem based on the structure of MANNER, the SOTA model of the current time domain.

· The objective function is defined in terms of both the time domain and the time-frequency domain, but we design Metric Loss for direct performance evaluation index improvement.

Encoder-Decoder Architecture Model and Anti-Aliasing

Speech enhancement studies are divided into time domain models and time-frequency models according to input and output. Until recently, time domain models have shown inferior performance compared to time-frequency domain models. However, time domain models such as DEMUCS[1], SE-Conformer[2], and MANNER[3], which have been proposed since 2021, have shown that the best performance can be achieved or close to it. MANNER is the best performing model of the time domain at the time of writing this proposal [3]. The overall structure of the MANNER is shown in Figure 1.

MANNER is composed of multiple encoders and decoders and is based on a U-Net structure with skip connections between the corresponding blocks. The encoder and decoder layer consists of three blocks: Down/Upconvolution, ResCon, and Multi-view Attention. Among them, Down convolution is used only in encoders and Up convolution is used only in decoders. Down convolution plays a role in reducing the input size by adjusting the stride of the convolution operation. Up Convolution uses Transposed Convolution to increase input size. Meanwhile, Multi-View Attention uses three operations simultaneously: Channel Attachment, Global Attention, and Local Attention. Among them, Max pooling and Average pooling are used in Channel Attention and Local Attention.

In addition, MANNER uses both time domain distance and time frequency domain distance functions as objective functions, even though they are time domain models. (Equations 1, 2)

Here, y denotes a noise-free voice, R denotes an STFT resolution, and T denotes a voice reproduction time. The aforementioned three models of DEMUCS (Figure 2), SE-Conformer (Figure 3), and MANNER (Figure 1) were all able to utilize the U-Net structure to produce good results. That is, based on the encoder-decoder structure, the skip connection between the corresponding encoder and the decoder was applied.

Models based on these U-Net structures are exposed to aliasing problems during the down/up sampling process. Zhang et al. [4]revealed the ease of aliasing in the process of max pooling and convolution pooling, which are widely used in general deep learning. Shi et al. [5] pointed out the aliasing problem in transposed convolution, a commonly used upscaling method. [4]and [5] proposed MaxBlurPooling (Figure 4) and sub-Pixel convolution (Figure 5) as solutions to aliasing.

Metric Loss Specific to Speech Enhancement Model

MetricGAN[6] proposed Metric Loss (Equation 3) which uses non-differentiable performance metrics as the objective function of neural networks.

(3)

Here, x represents a noisy voice, y represents a noise-free voice, G represents a voice improvement model, D represents a metric estimation model, and Q represents an evaluation index.

Using Metric Loss increases the correlation between model learning and performance improvement over existing L1 and L2 Loss [6]. Metric Loss showed significant performance improvement on various evaluation indicators, and studies such as CMGAN[7] and MetricGAN+[8] also adopted it as an objective function.

Anti-aliased Encoder-Decoder Architecture with Metric Loss

The proposed study is based on the model structure of MANNER, which shows the best performance among the aforementioned models, but aims to develop a model that is robust to aliasing. To this end, the techniques of blurpooling and sub-pixel convolution will be introduced. In addition, by introducing Metric Loss suitable for voice improvement, model training aims to achieve the qualitative excellence of voice improvement.

2) Research objectives and research contents for the second year

[Research goal for the 2nd year]

· We develop the best performance model of the time-frequency domain.

· To this end, we develop a speech enhancement model utilizing both Short Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT).

· The objective function is defined in terms of both the time domain and the time-frequency domain, but allows Metric Loss to be realized in various domains such as Magnitude and CWT results.

[2nd year's research]

The time-frequency model has time-frequency domain data converted by applying STFT to time domain data as input and output. Applying STFT to time domain data can obtain a complex number output. In the early days of the time-frequency model, only magnitude of the STFT results was used. Since then, PHASEN [15], a model that uses both Magnitude and Phase, has been published, considering that phase has a significant impact on real speech recognition. Later, FullSubNet+[16] used complex number data without conversion to polar coordinate system as input, while models such as DB-AIAT[17] and CMGAN[7] used complex number and magnitude as input.

Meanwhile, in the Lee et al. [10] study, DWT was applied to the magnitude image to subdivide the magnitude into aproximate and detail parts, and then process each part separately (the methods proposed by CMGAN[7], PCS[10], and TENET[18] were verified immediately after the Voice+Botate-EM data.

Notable key ideas in the time-frequency domain methodology are as follows. First, CMGAN[7] is a model developed based on MetricGAN[6] and Conformer[19]. MetricGAN proposed a method to replace performance evaluation indicators of non-differentiable speech enhancement with differentiable neural networks through GAN structures. Conformer[19] is a structure that has been mainly used in speech recognition in the past, and it is a model that has the advantages of both Transformer[20] which is advantageous for global characteristic extraction and CNN which is advantageous for regional characteristic extraction CMGAN consists of a Generator network and a Discriminator network. Based on the Encoder-Decoder structure, Generator generates improved speech by applying time-axis conformers and frequency-axis conformers sequentially between Encoder and Decoder. Discriminator estimates non-differentiable performance metrics. CMGAN achieved SOTA performance.

Second, PCS[9] is a methodology for improving a target sound source of voice improvement. The PCS applies gamma correction to the Spectorgram for voice improvement to convert the target sound source to a clearer voice. At this time, the correction range for each band was different based on the critical band importance for voice. The calculation method of PCS is shown in Equation 6.

Table 2 shows γ for each band. The PCS-processed target sound source improved the performance of various voice reinforcement models [9]. In addition, PCS can be used as a tool for post-processing of voice enhancement.

Third, TENET [18] proposed a method for effectively learning a voice enhancement model. TENET proposed a method of simultaneously learning a data pair in which the time of the input and target sound source was reversed and a data pair in which the inversion was not applied using a Siamese network (Figure 9). With this, TENET has achieved good results in public data sets.

CMGAN, PCS, TENET, Lee et al. Each core module proposed by the study is model independent. Therefore, a voice reinforcement model with improved performance can be developed through a model structure that appropriately utilizes each module.

3. Promotion strategy, method and promotion system of research project

1) Strategies and methods for promoting research projects

The conduct of this study consists of several experiments. Each experiment consists of three steps: implementation of the research goal model, data collection, and confirmation and verification of the experiment results. To implement the research target model, use PyTorch, an open source deep learning framework under development at Meta. PyTorch is a framework that is already being used in many studies, and supports GPU-accelerated computations such as CUDA, enabling high-speed model verification. The experimental data uses VoiceBank+DEMAND, which is public data. For experimental verification, evaluation indicators such as PESQ and STOI should be used. For fairness, it is measured using python libraries, pypesq and pystoi, respectively.

2) Promotion system of research projects

- Participated by 1 research director, 2 doctoral students, and 2 master's students

3) Research period and research fund appropriateness

- Research expenses are calculated based on labor costs (participation rate of 30% or more) for 2 doctoral students and 2 master’s students

- Deep learning equipment is owned by the laboratory, so there is no need to purchase or rent additional equipment. Currently, there are 2 RTX A6000 units, 4 TITAN RTX units, and multiple RTX-3090 labs

- Calculation of research activity expenses for participation in domestic and foreign academic societies and calculation of thesis publication costs (2 cases each of domestic and foreign thesis presentations per year, 1-2 SCI-level thesis publications)

4. Researcher's ability to conduct research

- Published 33 domestic and international academic papers in the field of machine learning and computer vision

- Registered about 30 patents for employee inventions and has numerous technology transfers

- Commendation from the Minister of Small and Medium Venture Business (2021), Korea Industry-Academia-Research Association President’s Award (2018), University President’s Award for Excellence in Industry-Academy Cooperation (2017)

- Conducted its own preceding research on voice enhancement and presented related thesis [23] at the conference (May 2022)

- Possession of practical technology development capabilities through various industry-university cooperation technology tasks in the field of computer vision

5. Utilization plan and expected effect of research project

1) How to use the research project

The speech enhancement model is widely used as a preprocessing process for speech recognition models. Since this study aims to reach the performance of the latest model with the highest performance, the results of this study will help improve the recognition performance of the speech recognition model. In addition, the voice enhancement model can be used for preprocessing purposes of assistive devices for the hearing impaired. The purpose of injection of hearing aids is to assist users in recognizing their voices. Most hearing aids amplify sound ranges that are difficult for the user to hear through an external microphone and ear speaker. At this time, if the voice enhancement of this study is applied to the microphone input, it will be more helpful for the hearing impaired to recognize the voice.