**Department of Electrical, Computer, and Software Engineering**

**Part IV Research Project**

Final Report

Project Number: 58

The reconstruction of speech/voice with the use of noise-cancelling algorithms and machine learning

Edward Chan

Timothy Aguana Cabrera

Catherine Watson

Yusuke Hioka

13/10/2023

**Declaration of Originality**

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

A close-up of a signature

Description automatically generated

Name: Edward Chan

ABSTRACT: A longstanding problem in audio signal processing is the speech issue in noisy environments when the signal-to-noise ratio (SNR) is low, resulting in significantly quieter speech than in higher SNR. Speech enhancing and separation methods exist using statistical techniques or machine learning algorithms to solve the low SNR problem. In order to overcome this problem, we propose an iterative learning approach for speech enhancement based on two machine learning algorithms that generate and classify a noisy mixture. The algorithm begins with an end-to-end learning method based on a generative adversarial network (GAN) that uses noise estimation and spectral attenuation to remove the noise of the noisy mixture to create estimated clean speech. At the same time, another learning method based on the convolutional neural network (CNN) uses a spectrogram representation of the estimated clean speech to distinguish between clean and noisy speech by classifying SNR levels. The question that the paper attempts to answer is if the SNR level of a noisy mixture can represent an improved quality or intelligibility in determining between clean and noisy speech. Our proposed method will answer this question using computational matrices of perceptual evaluation of speech (PESQ), short-time objective intelligibility (STOI), and signal distortion ratio (SDR) to validate the algorithm.

**Contents**

[ABSTRACT: 3](#_Toc147852920)

[List of Figures 6](#_Toc147852921)

[Abbreviations 7](#_Toc147852922)

[1. Introduction 8](#_Toc147852923)

[2. Literature Review 9](#_Toc147852924)

[2.1. Speech Enhancement and Speech Separation 9](#_Toc147852925)

[2.2. Time-Frequency Domain 10](#_Toc147852926)

[2.3. Learning Machines 11](#_Toc147852927)

[2.3.1. Multilayer perceptron (MLP) 11](#_Toc147852928)

[2.3.2. Convolutional neural network (CNN) 12](#_Toc147852929)

[2.3.3. Generative adversarial networks (GAN) 12](#_Toc147852930)

[3. Research Method 13](#_Toc147852931)

[3.1. Datasets 14](#_Toc147852932)

[3.2. Experiment Setup 14](#_Toc147852933)

[3.2.1. Spectrograms 14](#_Toc147852934)

[3.2.2. CNN Classification Model Parameters 16](#_Toc147852935)

[4. Experiment and Results 17](#_Toc147852936)

[4.1. Experimentation of CNN Classification Model 17](#_Toc147852937)

[4.1.1. Implementation Details 17](#_Toc147852938)

[4.1.2. Results of the CNN Classification Model 17](#_Toc147852939)

[4.2. Evaluation Methods 19](#_Toc147852940)

[4.3. Evaluation results based on computational measurements 19](#_Toc147852941)

[5. Discussion 22](#_Toc147852942)

[6. Conclusion 24](#_Toc147852943)

[Acknowledgements 25](#_Toc147852944)

[References 25](#_Toc147852945)

# List of Figures

[Fig. 1. Processing scheme for single channel and multichannel speech enhancement and separation. 10](#_Toc147852191)

[Fig. 2. High-level flow chart of the proposed design. 13](#_Toc147852192)

[Fig. 3. Wideband spectrogram representation. 15](#_Toc147852193)

[Fig. 4. Narrowband spectrogram representation. 16](#_Toc147852194)

[Fig. 5. Overview of the CNN classification model with each layer. 16](#_Toc147852195)

[Fig. 6. CNN classification model training accuracy at each epoch. 18](#_Toc147852196)

[Fig. 7. Confusion matrix of the testing of the CNN classification model. 18](#_Toc147852197)

[Fig. 8. Babble noise for (a) PESQ score, (b) STOI score, and (c) SDR score. 20](#_Toc147852198)

[Fig. 9. Train station noise for (a) PESQ score, (b) STOI score, and (c) SDR score. 21](#_Toc147852199)

# Abbreviations

**SNR S**ignal-to-**N**oise **R**atio

**PESQ P**erceptual **E**valuation of **S**peech **Q**uality

**STOI S**hort-**T**ime **O**bjective **I**ntelligibility

**SDR S**ignal **D**istortion **R**atio

**MLP M**ulti-**L**ayer **P**erception

**CNN C**onvolutional **N**eural **N**etwork

**GAN G**enerative **A**dversarial **N**etwork

**DEMAND D**iverse **E**nvironmental **M**ultichannel **A**coustic **N**oise **D**atabase

**ReLU R**ectified **L**inear **U**nit

# Introduction

Speech enhancement refers to the problem of improving the quality of speech in terms of intelligibility or perceived quality using enhancing programs to determine whether the source extracted is a human voice or background noise. Techniques for enhancing speech are used in many ways, such as transforming a raw audio signal into its spectral representation to determine the difference between the noise signals and the speech signals. With spectral representations, a newly introduced machine learning method can be utilised to manipulate the noisy additive signal and subtract it from the raw audio signal to output a cleaner speech signal.

Extensive research on machine learning methods has claimed to work better than original solutions of well-known statistical interference techniques such as Wiener filtering [1] and the Gaussian mixture model [2]. Most of them test the different machine learning methods using a similar setup of an addictive dataset of clean speech and noise with varying signal-to-noise ratio SNR levels [1, 3, 4] to compare and evaluate the results using the original solutions, which are limited in their performance at lower SNR levels. Nevertheless, none use SNR as a form of class to determine if the audio signal has a louder speech signal than the background noise, which can be represented as a method to determine if the audio signal is intelligible.

This paper investigates the difference between speech enhancement and speech separation, the time-frequency domain, existing machine-learning approaches and the problem with existing speech enhancement methods in section 2. Section 3 discusses the overall research method and the steps for accomplishing the project, while section 4 contains the measures taken for the experiment and the results obtained. Section 5 discusses what was found, the experimentation's limitations and the validation of the results. Finally, section 4 highlights relevant contributions with a final summary and future considerations.

# Literature Review

## Speech Enhancement and Speech Separation

Speech enhancement is the segregation of speech and suppression of all interfering background noises used to improve speech quality and intelligibility [5]. Speech separation, or source separation, separates one or more target speech signals while suppressing interfering sources and noise [5, 6, 7]. Besides speech, source separation is also necessary for separating music and movie soundtracks. In music, there are usually one or more instruments playing together with a singer's voice that provides lyrics, while for movies, there is usually speech with background noise and sound effects. Both these audio processing terms are often interchangeable as they both refer to the suppression of interfering speakers and noise from a speech signal. For both practices, the goal is to solve either the high or low SNR issue. For example, professional music tracks typically have a high SNR because the singer's voice is usually louder than the background instruments and sound effects, which makes the signal clear and easy to understand. When listening to a radio with a weak signal, characterised by a lot of static and interference, the SNR is low as the background noise overpowers the signal, making the audio distorted or unintelligible to the listener. The number of output microphones, also considered I channels where I am represented as , is an important consideration as a microphone with I =1 channel is called a single channel and is represented by a scalar. Moreover, a signal with I > 1 channel is called multichannel and is represented by a vector [5].

The general approach for speech enhancement and separation processing uses single-channel or multichannel scenarios, as depicted in Fig. 1 [5]. Beginning with a mixture signal of the target and noise source in the time domain , which is transformed into its frequency domain using a short-time Fourier transform to obtain a magnitude and phase spectrogram representation, a visual representation of the audio signal. The spectrogram can also be derived from its frequency representation over a discrete time interval, considered the time-frequency domain . Prior information about the scenario is utilised to aid the parameter estimation. Given the parameters, the spatial or spectral filtering which performs image processing from a chosen speech modelling approach is derived and applied to the mixture to create an estimated source. for single channel or source spatial image for multichannel. Lastly, we perform the inverse of the frequency transform to obtain the time domain signal, yielding time domain source estimates. or source spatial image signal .



Fig. 1. Processing scheme for single channel and multichannel speech enhancement and separation [5].

## Time-Frequency Domain

Working with the raw audio signal in the time domain is not standard for audio signal processing. Instead, most methods transform the inputs to their time-frequency representation to model speech enhancement or separation [3]. The reasons why the time-frequency domain is used instead of the time or frequency domain are as follows. The sound source obtains mixtures of noise and speech, which cannot be approximated with just the amplitude of sound in the time domain, nor can the frequency spectrum represent it in the frequency domain. The mixed sound source is better defined in the time-frequency domain as the amplitude of a sound as a function of both time and frequency jointly accounts for its temporal and spectral characteristics [5]. Lastly, a property of the time-frequency domain is that sound sources are more sparsely distributed and overlap less, resulting in an easier separation. A short-time Fourier transform converts a signal from the time domain to the frequency-time domain and back. The short-time Fourier transform technique allows for audio source segmentation, where the signal is divided into short, overlapping frames multiplied with a window function to avoid spectral leakage [5]. The windowed frames are then placed into the Fourier transform to produce spectrograms, 2D representations of a signal frequency content over time. A magnitude represents the spectrogram, the colour intensity of the frequency content at different points in time, and a phase supplies information about the spectral component at other points in time. However, most single-channel audio processing methods do not use phase, as the frequency component of a source is not affected by the phase but instead affected by the room impulse response. Room reverberation and spatial sources can affect the phase in multichannel audio processing [4].

## Learning Machines

Most of the literature uses supervised learning to train algorithms by providing a noisy sound source and a clean speech source or noise source. Unlike the unsupervised learning method, which contains no clean speech or noise sources knowledge, the training algorithm relies on specific instructions or clusters within the data. Some deep neural network algorithms used for supervised speech enhancement and separation include multilayer perceptron (MLPs), convolutional neural networks (CNNs), and generative adversarial networks (GANs).

### Multilayer perceptron (MLP)

MLP with a feedforward connection using one hidden layer is the most popular neural network and is trained using an error backpropagation algorithm where the network weights are adjusted to minimise the prediction error through gradient descent [6]. The predicted error is the difference between the predicted and actual output. The MLP algorithm only uses one hidden layer due to the vanishing gradient problem caused by adding more layers, resulting in little weight change that impacts biasing.

### Convolutional neural network (CNN)

People use CNN for pattern recognition, object detection and other computer vision tasks. However, convolutional filters used to extract features from the noisy input signal to output an enhanced signal that removes the noise are also another popular neural network to improve signal processing. A U-Net CNN is an architecture based on CNN that researchers introduced in biomedical imaging. It is trendy in cell segmentation, organ segmentation, and other medical image analysis applications.

U-Net consists of an encoder-decoder structure with skip connections to capture fine-grained spatial information. The architecture can then successfully perform image analysis tasks by generating segmentation maps. Researchers have adapted this method to other fields where image segmentation is required, such as speech enhancement [8].

### Generative adversarial networks (GAN)

Two models are trained simultaneously in GAN; one of the models is the generator that learns to create the noisy speech samples into a clean speech sample. While the discriminator, usually a binary classifier, learns to distinguish between noisy and clean speech samples. This concept involves a game-like setting where people train the generator to fool the discriminator while the discriminator tries to determine if someone is fooling it [6].

# Research Method

Reconstruction of speech from a noisy, unintelligible source can be used in various applications such as voice communication, hearing aid devices, speech recognition, and telecommunication devices [1, 6]. The problem that the research seeks to solve is the issue of speech at low SNR, where the source signal is hard to distinguish between in a noisy mixture. Current speech enhancement methods struggle to improve speech quality while preserving speech intelligibility in a non-stationary noise environment [1].

A diagram of a diagram

Description automatically generated

Fig. 2. High-level flow chart of the proposed design.

In Fig. 2, our high-level diagram shows the training and testing requirements. Starting the training of the overall system, we train the GAN model using a range of noises that will be further discussed to quantify the energy associated with each frequency component within the noise. This comprehensive analysis of frequency components will allow us to fine-tune the GAN architecture and training parameters to ensure that the generated data accurately captures a more accurate noise power spectral density called the codebook. As mentioned, the generator predicts the codebook when training the GAN model. At the same time, the discriminator aims to distinguish between a less accurate and more accurate codebook, which updates the generator to enhance its prediction accuracy. Testing the system includes a noisy mixture of speech at any SNR level through the GAN model to generate an estimation of the codebook. Based on the noisy mixture, the adaptive gain and filter adjust the frequency amplitude and response based on the input signal. A speech estimation is then generated and critiqued using a CNN classification model to distinguish between different SNR levels for clean or noisy speech. If the critic determines that the generated mixture does not meet the criteria for clean speech, the system initiates a regenerative process. During this process, a modified adaptive gain adjustment control is applied to the mixture signal and continuously iterated until one of the two conditions is met. These conditions are if the system recognises clean speech when the power of the speech signal substantially exceeds the background noise power signal, resulting in high SNR or when the system reaches a predefined threshold.

## Datasets

The dataset used to train the models is generated from two sources. The speech data comes from the Voice Bank corpus [9], which contains 500 native English speakers who read about 400 sentences, containing more than 300 hours of recordings. The environmental background noise comes from the Diverse Environment Multichannel Acoustic Noise Database (DEMAND) [13]. It includes 13 different real-world noises 5 minutes long in various environmental situations such as on a bus, a café or an office. For the CNN classification model, 1,200 utterances were used for training, 100 for validation and 100 for testing, and all utterances were randomly combined with 1 of the 13 different noise classes to create a noisy mixture. The recording studio quality for the Voice Bank corpus was sampled at 48kHz, but we downsampled the quality sample to 16kHz for this study to match the DEMAND quality sample. For both training and testing, all the samples were mixed at seven different SNRs: 9, 6, 3, 0, -3, -6, -9dB and a clean speech without any noise mixture. This results in 9,600 training samples and 1,200 validation and testing samples from unseen speakers. The samples are, on average, 5 seconds long and do not use preprocessing, such as voice activation or filtering.

## Experiment Setup

### Spectrograms

The proposed dataset is transformed into its wideband spectrogram representation with Short Time Fourier Transform parameters [11] of N-FFT = 1024, window function = Hann, hop length = 80 samples and sample per window = 320 samples. Using a wideband rather than a narrowband spectrogram would result in a better representation of differentiation between speech and noise within the spectrogram, as the acoustic features are better represented. This is evident in Fig. 3, which is the wideband, and Fig. 4 is the narrowband spectrograms, as the former shows more apparent individual pitch harmonics resolved in frequency than the latter [16].

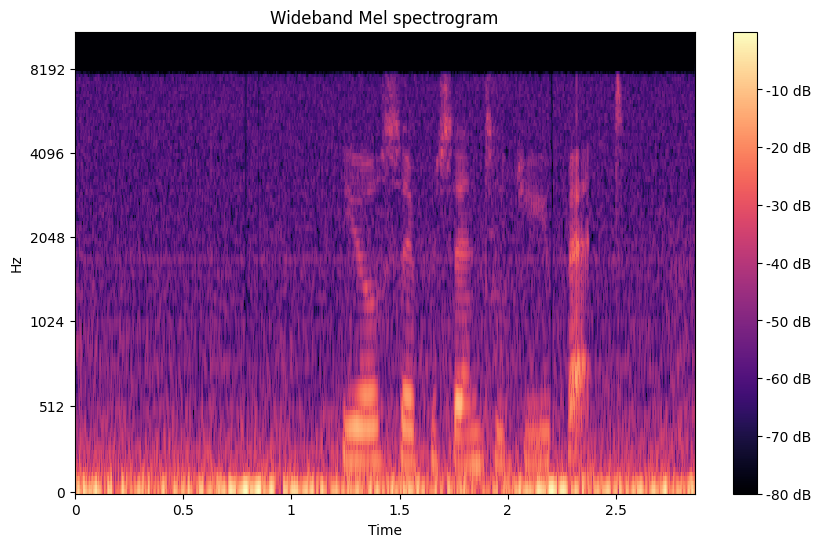


Fig. 3. Wideband spectrogram representation.

A close-up of a screen

Description automatically generated

Fig. 4. Narrowband spectrogram representation.

### CNN Classification Model Parameters

The proposed model in Fig. 5 includes convolution layers to extract features from the spectrogram and pooling layers to downsize the image so features can be detected at various resolutions [12]. Using both the convolution and pooling layers for feature extraction, the next step is to use a flattening layer that converts the 2D features from the convolutional layers into a 1D vector to connect the convolutional layers to a fully connected layer. The first dense layer has 1024 neurons with ReLU activation, which tries to understand the complex relationship in the patterns. The final dense layer has eight neurons with softmax activation, which outputs probabilities for each SNR class.

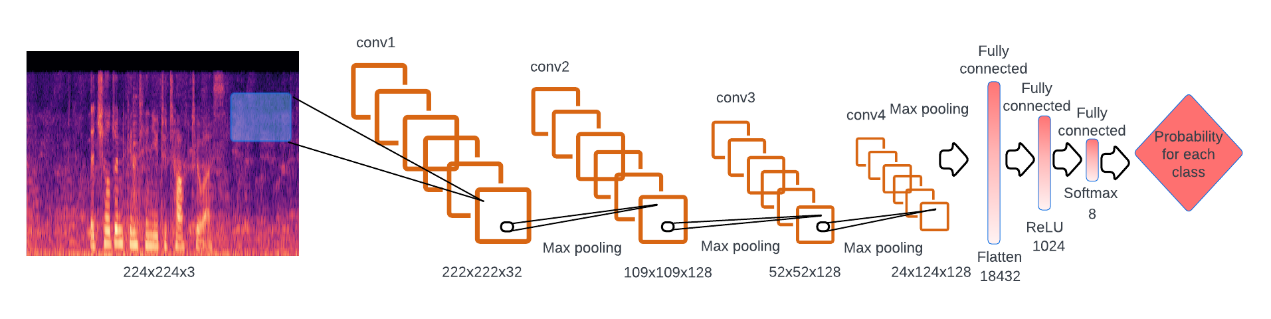


Fig. 5. Overview of the CNN classification model with each layer [12].

# Experiment and Results

## Experimentation of CNN Classification Model

### Implementation Details

Distinguishing between whether speech is considered clean or not based on the audio signals SNR level should mean that the speaker's voice is louder than the background noise source, therefore resulting in speech that should be intelligible.

For this experiment, the class considered clean speech when the signal exceeds the background noise, 9, 6 and 3dB SNR levels. SNR 0, -3, -6 and -9dB will be considered noisy speech because the background noise power is much greater than the speaker's source signal.

### Results of the CNN Classification Model

Training the model with 80% of the total dataset using the conventional neural network with a batch size of 10 and epoch of 20 resulted in an overall training accuracy of approximately 97%, as shown in Fig. 6. Also seen in Fig. 6 is the accuracy of the validation dataset at each epoch in the model with 10% of the total dataset. At around 12.5 epochs, the validation accuracy is 90%, which results in minimal change to accuracy at each new epoch and has reached a point of diminishing returns in learning from the same dataset. In Fig. 7, a confusion matrix of testing the model with the remaining 10% of the dataset is used to determine if the model is accurate with unseen data. The testing accuracy is 90.25%, with most incorrect labels varying by an increase or decrease of 3 decibels.

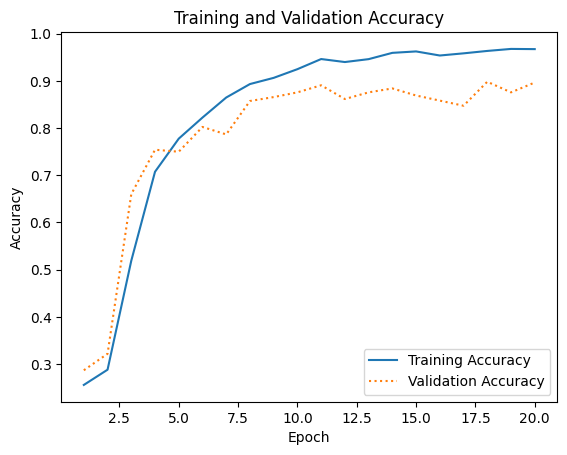


Fig. 6. CNN classification model training accuracy at each epoch.

A graph of data with numbers and symbols

Description automatically generated with medium confidence

Fig. 7. Confusion matrix of the testing of the CNN classification model.

## Evaluation Methods

To evaluate the quality of the enhanced speech, we used the objective metric of perceptual evaluation of speech quality (PESQ) [13] and, for intelligibility, the short-time objective intelligibility score (STOI) [15]. In comparison, signal distortion ratio (SDR) [14] is used to evaluate both the quality and intelligibility. The range of the PESQ score is from -0.5 to 4.5, which indicates better speech quality with higher values. The range of SDR score is expressed in decibels and has a wide range depending on the audio signal, with higher values indicating better signal quality and intelligibility. In contrast, the range of STOI score is from 0 to 1, which expresses the percentage of intelligibility in the signal.

## Evaluation results based on computational measurements

We compared our approach using two different noise types: babble noise, the background noise of people talking in Fig. 8 and train station noise, which is disruptive noise generated by trains and background ambience in Fig. 9. All three objective metrics are compared with the proposed method of the GAN model generation of speech-enhanced audio using the classification model to criticise the enhanced audio based on its SNR levels. The metrics are also compared to the GAN model generation of speech-enhanced audio without the criticism of the classification model and instead based on the highest metric score for each metric. Both these measurements used an alpha value, a parameter in spectral subtraction, to control the balance between noise reduction and speech preservation. When the alpha value is small, there is less aggressive noise suppression, while a more significant alpha value will result in more aggressive noise suppression. The whole system goes through a maximum of 40 iterations where the alpha step size is 0.05, and the maximum alpha threshold value is 2, which results in an unintelligible speech at any higher alpha value. However, if the classification model finds the audio is enhanced to an SNR level equal to or greater than 3dB, the optimal audio is found, and iterations stop. Lastly, the metric is compared to the original audio before enhancement to show if the method made a difference in computational measurements.

A screenshot of a graph

Description automatically generated

Fig. 8. Babble noise for (a) PESQ score, (b) STOI score, and (c) SDR score.

A screenshot of a graph

Description automatically generated

Fig. 9. Train station noise for (a) PESQ score, (b) STOI score, and (c) SDR score.

# Discussion

For all cases in Fig. 8 and Fig. 9, the proposed algorithm does not improve speech when the SNR is equal to or higher than 3dB, as the methodology considers this speech already intelligible. For the proposed method with and without the CNN classification model, we can see minimal change in audio quality regarding the PESQ score for both the babble noise Fig. 8 (a) and train station noise Fig. 9 (a). For the train station noise, the proposed method with only GAN has an increase in audio quality of 0.01 at -9dB, which shows that quality improvement can be found at lower SNR levels. The proposed method resulted in a slightly worse intelligibility score than the original audio signal at lower SNR for the STOI score for both the babble noise Fig. 8 (b) and train station noise Fig. 9 (b). In contrast, the proposed method with only the GAN that takes the optimal score at each iteration gave slightly better intelligibility at lower SNR than the original audio signal. The SDR score, which focuses on speech quality and intelligibility of the entire audio signal, scored for both the proposed methods with and without the classification model, gives significant separation performance with the babble noise Fig. 8 (c), resulting in approximately 1.5 score increase. While the train station noise, Fig. 9 (c), portrays an increase in 4 while increasing at each decreasing SNR value.

One problem that may limit the performance of the CNN classification is that there is no class for unintelligible speech like the STOI [13] and PESQ [15] objective matrices; therefore, if the proposed method gets fed just noise without speech, it will keep attempting to find the speech until the threshold is reached. Another possibility is that the classification model is trained on the spectrogram noise instead of the spectrogram speech as the optimal evaluation matrices are reached before the model determines the speech as high SNR, also known as the enhanced speech.

Overall, the proposed GAN model method gives slightly better quality and intelligibility. In most cases, the CNN model considers the audio signal to have a high enough SNR level to be considered enhanced speech after reaching the optimal PESQ and STOI score values. However, the SDR score improves results as the SNR level decreases, which is evident with the decrease in the bass in the audio from informal listening tests to confirm these results.

# Conclusion

This paper investigates a novel approach to improve noise generalisation by cycling a low SNR signal through a learning machine until it is considered a high SNR signal, considered enhanced speech. Using the speech enhancement algorithm, we showed that the speech contaminated with babble and train station noises at different SNR levels slightly improved intelligibility and quality. However, it is overcycling through the CNN classification model. These results indicate that although the formal software evaluation tools may not directly correlate to SNR improvements with increased intelligibility or quality, informal listening suggests improved perceived quality. This improvement is likely due to the reduction in bass from the noisy background signal as the SNR increases.

Future works will include the addition of the class unintelligible speech in the classification model and further research that speech signals may be lost with low SNR, a predictive model could be used to reconstruct lost signals. Furthermore, training the classification and GAN model with a wide range of realistic and simulated noise data can further generalise the model's ability to handle different noise inputs. Additionally, having access to better computational power and the ability to train models for longer durations can significantly enhance the performance and capabilities of the models.

# Acknowledgements

We want to thank the supervisor, Catherine Watson, and co-supervisor, Yusuke Hioka, for providing the support and guidelines to make this project successful.

# References

1. Manamperi, W., Samarasinghe, P. N., Abhayapala, T. D., & Zhang, J. (2022). GMM Based Multi-Stage Wiener Filtering for Low SNR Speech Enhancement. *International Workshop on Acoustic Signal Enhancement, IWAENC 2022 - Proceedings*, 0–4. https://doi.org/10.1109/IWAENC53105.2022.9914707
2. Delcroix, M., Kubo, Y., Nakatani, T., & Nakamura, A. (2013). Is speech enhancement preprocessing still relevant when using deep neural networks for acoustic modeling ?, (August), 2992–2996. https://www.isca-speech.org/archive/pdfs/interspeech\_2013/delcroix13\_interspeech.pdf
3. Rethage, D., Pons, J., & Serra, X. (2018). A wavenet for speech denoising. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018*-*April*, 5069–5073. https://doi.org/10.1109/ICASSP.2018.8462417
4. Saleem, N., Irfan Khattak, M., Ali, M. Y., & Shafi, M. (2019). Deep neural network for supervised single-channel speech enhancement. *Archives of Acoustics*, *44*(1), 3–12. https://doi.org/10.24425/aoa.2019.126347
5. Vincent, E., Virtanen, T., & Gannot*, S. (2018). Audio source separation and speech enhancement. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119279860*
6. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *26*(10), 1702–1726. https://doi.org/10.1109/TASLP.2018.2842159
7. Zhang, C., Hui, C. T. J., Hioka, Y., & Watson, C. I. (2022). Performance Of Speech Enhancement Algorithms For Native Mandarin Listeners On English Perception. *International Congress on Acoustics*, 1–6. https://guarant.cz/icphs2023/360.pdf
8. Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks. *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, 745–751.

https://ejhumphrey.com/assets/pdf/jansson2017singing.pdf

1. Veaux, C., Yamagishi, J., & King, S. (2013). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*. https://doi.org/10.1109/ICSDA.2013.6709856
2. Thiemann, J., Ito, N., & Vincent, E. (2013). The Diverse Environments Multichannel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. *Proceedings of Meetings on Acoustics*, *19*, 1–6. https://doi.org/10.1121/1.4799597
3. *librosa.feature.melspectrogram — librosa 0.10.1 documentation*. (2013). https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html
4. *Deep-Learning/Audio Classification (CNN).ipynb at master · jeffprosise/Deep-Learning · GitHub*. (n.d.). Retrieved October 10, 2023, from https://github.com/jeffprosise/Deep-Learning/blob/master/Audio%20Classification%20(CNN).ipynb
5. A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 2001, pp. 749-752 vol.2, doi: 10.1109/ICASSP.2001.941023.
6. Villamin, F. J., & Prado, D. (2016). Design and Implementation of Speech Enhancement System to make Rapid Voice Development robust to noise. https://researchspace.auckland.ac.nz/handle/2292/29155
7. C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, Sept. 2011, doi: 10.1109/TASL.2011.2114881.
8. A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," in IEEE Spectrum, vol. 7, no. 8, pp. 57-62, Aug. 1970, doi: 10.1109/MSPEC.1970.5213512.