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

**Part IV Research Project**

Final Report

Project Number: 58

Low-SNR speech enhancement through GANs and codebooks.

Timothy Aguana Cabrera

Edward Chan

Catherine Watson, Yusuke Hioka

[Date including year 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.

Name: Timothy Aguana Cabrera

# **Abstract**

This contribution aims to improve the denoising performance of codebook-based speech enhancement algorithms given against speech signals made unintelligible due to the high presence of noise. A proposed joint approach using codebooks derived from Gaussian mixture models (GMMs) and a Wasserstein GAN (WGAN) was developed to generate time-varying spectral candidates that closely resemble the underlying noise and speech spectra. The result of this contribution shows that the proposed approach is fully capable of accurately estimating the spectra of noise but fails in accurately estimating the spectra of speech, resulting in the distorted and unintelligible Wiener-filtered low signal-to-noise mixture.

# **Acknowledgements**

The student would like to thank the supervisor Catherine Watson and co-supervisor Yusuke Hioka

for providing valuable insights and support to make this project successful.

# **Table of contents**

[**Abstract** 1](#_Toc148116338)

[**Acknowledgements** 2](#_Toc148116339)

[**Table of contents** 3](#_Toc148116340)

[**List of Figures** 4](#_Toc148116341)

[**List of Tables** 5](#_Toc148116342)

[**List of Acronyms** 6](#_Toc148116343)

[**1. Introduction** 7](#_Toc148116344)

[**2. Literature review** 8](#_Toc148116345)

[**2.1.****Statistical estimators and neural networks.** 8](#_Toc148116346)

[**2.2. Generative adversarial networks and codebooks.** 9](#_Toc148116347)

[**2.3. Research intent** 10](#_Toc148116348)

[**3. The proposed approach** 12](#_Toc148116349)

[**3.1.** **Pre-existing works and the proposed approach.** 12](#_Toc148116350)

[**3.2. Derivation of codebooks from GMMs** 13](#_Toc148116351)

[**3.3. Generative adversarial networks** 14](#_Toc148116352)

[**4. Experiments & results** 17](#_Toc148116353)

[**4.1. Experimental setup** 17](#_Toc148116354)

[**4.2. Results & Discussion** 18](#_Toc148116355)

[*4.2.1. Wiener filtering* 18](#_Toc148116356)

[*4.2.2. Spectral subtraction* 22](#_Toc148116357)

[*4.2.3. Reflection on research objectives* 23](#_Toc148116358)

[*4.2.4. Research limitations* 24](#_Toc148116359)

[**5. Conclusion** 25](#_Toc148116360)

[**References** 26](#_Toc148116361)

# **List of Figures**

[Fig. 1. Proposed WGAN architecture. 16](#_Toc148114984)

[Fig. 2. Estimated spectrum of noise compared against the ground truth. 18](#_Toc148114985)

[Fig. 3. Spectrograms of an unfiltered, WGAN, and optimally Wiener-filtered -9dB waterfall noise audio file. 19](#_Toc148114986)

[Fig. 4. Spectrograms of an unfiltered, WGAN, and optimally Wiener-filtered -9dB babble noise audio file. 19](#_Toc148114987)

[Fig. 5. WGAN estimated PSD of speech against the ground truth. 20](#_Toc148114988)

[Fig. 6. Wiener-filtered babble mixtures with SNRs of -3 dB, -6 dB, and -9 dB (left to right) using the GAN estimated speech and noise PSD. 21](#_Toc148114989)

[Fig. 7. Filtered waterfall noise mixture with an SNR of -9dB and an incrementing alpha value (left to right) using the GAN estimated noise PSD. 23](#_Toc148114990)

# **List of Tables**

[Table 1: PESQ, STOI, and SDR scores of the semi-optimal Wiener-filtered waterfall audio file. 19](#_Toc148114991)

[Table 2: PESQ, STOI, and SDR scores of the semi-optimal Wiener-filtered babble audio. 20](#_Toc148114992)

[Table 3: PESQ, STOI, and SDR scores of the Wiener filtered-babble audio file. 21](#_Toc148114993)

[Table 4: PESQ, STOI, and SDR scores of the filtered waterfall audio file. 23](#_Toc148114994)

# **List of Acronyms**

CNN – Convolutional Neural Network

DNN – Deep Neural Network

FFT – Fast Fourier Transform

FGLA – Fast Griffin-Lim Algorithm

GAN – Generative Adversarial Network

GB – Gigabyte

GMM – Gaussian Mixture Model

MMSE-STSA – Minimum Mean Square Error Short-time Spectral Amplitude

PESQ – Perceptual Evaluation of Speech Quality

PSD – Power spectral density

RAM – Random Access Memory

SDR – Signal-to-Distortion Ratio

SNR – Signal-to-Noise Ratio

STOI – Short-Time Objective Intelligibility

VAD – Voice Activity Detector

WGAN – Wasserstein Generative Adversarial Network

# **1. Introduction**

Speech enhancement studies signal processing techniques and algorithms to remove unwanted noise from a corrupted speech signal. This contribution aims to improve the performance of single-channel speech enhancement approaches towards low signal-to-noise ratio (SNR) mixtures, as standard communication devices typically do not have access to multi-channel configurations that offer superior filtering performance [1]. Current approaches consist of model-based statistical estimators with realizable filters [2-6] characteristics and neural networks that perform non-linear and non-analytic filtering [7-9]. Although neural networks have improved over traditional statistical estimators at relatively high SNRs, the effective enhancement of low SNR mixtures remains challenging for both approaches. Therefore, this contribution aims to investigate if a hybrid approach between a generative adversarial network (GAN) and the widely used estimation method of codebooks will improve the intelligibility of enhanced low SNR speech signals and provide insight into the filter characteristics of the neural network.

This report presents an overview of the report layout. Section 2 presents a literature review to address the gaps and present the motivation of this research. Section 3 contains information regarding the paper this contribution was based on and the proposed GAN and codebook model. Section 4 presents the results and discussion regarding the speech enhancement performance of the proposed model. Lastly, Section 5 presents the conclusions based on the findings of this research, contributions, and potential future work.

# **2. Literature review**

## **2.1.****Statistical estimators and neural networks.**

In many recent SCSE implementations, deep neural networks (DNNs) have consistently outperformed statistical filters in objective and subjective measures [10, 11]. The estimation of noise spectra is commonly done through some form of a statistical model that captures an expected distribution of the signal based on its characteristics. These statistical models typically use assumptions based on the signal, such as statistical stationarity and noise additivity, to determine spectral gain functions that result in an estimated clean signal. An example of a statistical estimator that utilizes these assumptions is the minimum mean square error short-time spectral amplitude estimator (MMSE-STSA). This estimator derives an optimum spectral amplitude gain function based on speech presence (or absence) and assumes that the statistical nature of the noise is stationary [3]. However, the performance of this estimator decreases substantially with low SNR mixtures due to the non-stationarity or high presence of noise compared to speech, resulting in residual noise or unintelligible speech caused by introduced distortions or over-attenuation [2]. Other works have proposed using Gaussian mixture models (GMMs) to obtain better noise estimates depending on the observed mixture and the approximated multimodal distribution. However, the resulting model may suffer from the initial assumption on the chosen number of modes and the assumption that each mode is of a Gaussian distribution [4-6].

In contrast, since its introduction to speech enhancement, neural network approaches have consistently outperformed statistical estimators at low SNR speech enhancement [7, 8]. Unlike statistical estimators, neural networks do not use assumptions and can learn complex multimodal distributions and non-linear relationships. Typical applications of these neural networks consist of 'end-to-end' time-domain signal synthesis and time-frequency masking-based approaches from noisy spectrograms [8, 9]. However, as addressed in the study [12] conducted by Nossier et al., the performance of these algorithms is primarily attributed to the chosen architecture, specified hyperparameters, and training data, where the direct relationship between these three factors on the speech enhancement performance of DNNs remains unclear in contrast to statistical estimators. In addition, the results from a study regarding informed source separation [13] show that statistical estimators can surpass the performance of neural networks if data availability is sparse. Therefore, this highlights the requirement of neural networks to be trained on extensive datasets to be generalizable in contrast to statistical estimators that rely on a limited pre-defined number of spectral candidates to estimate noise or speech spectra. However, with the increasing amount of public access to noise and speech datasets, the effects of this limitation are less prevalent.

## **2.2. Generative adversarial networks and codebooks.**

The GAN is a neural network that utilizes the discriminative nature of the discriminator neural network to train a generator neural network to generate plausible data iteratively. A typical application of GANs is the generation of realistic images through an input feature vector derived from random noise or feature extraction algorithms that allow for random or controlled generations, respectively [14, 15]. However, in recent years, GANs have also been applied to speech enhancement utilizing the advancements in image processing capabilities of convolutional neural networks (CNNs) to generate time-frequency magnitude masks from an input spectrogram or direct time-domain signal synthesis from a mixture that is learned through an auto-encoder mechanism [16, 17]. Although the mentioned works have achieved good denoising performance while enhancing speech quality or intelligibility at SNRs above 0dB, low SNR speech enhancement remains a challenge for GAN-based speech enhancement algorithms. In addition, current speech enhancement GANs typically need large training sets due to the high dimensionality of the desired functionality and often result in non-analytic filter characteristics. As a potential alternative, a well-known derivation of the required Wiener filter inputs uses codebooks containing potential noise or speech spectrum candidates. Typical codebook approaches require vast amounts of entries to accurately estimate various noise and speech spectra at the cost of computational efficiency [18-20]. As an alternative, a study [21] conducted by Chehresa and Savoji showed that formulating an overdetermined system made it possible to estimate the spectra of both noise and speech from only a limited number of spectral candidates and parameters. The proposed model was further developed using a multi-stage Wiener filtering approach with improved performance in enhancing low SNR mixtures of up to -10dB with non-stationary noise [22]. However, with the reduction in potential spectral candidates, it can be inferred that the proposed model will be sensitive to the chosen codebooks or spectral candidates and may potentially result in sub-optimal performance with certain mixtures. Therefore, to potentially improve on the proposed model, a GAN could be trained to generate optimal codebooks based on an observed mixture predictively. Current studies of neural network integration to codebook-driven methods are limited to spectral candidate selection [23], and through an extensive search, no studies regarding the utilization of GANs to predictively generate codebooks have been found. Although this is an understudied field in speech enhancement, it could hold promising results as it will give insight into the filter characteristics of the neural network through the generated codebook while potentially having lower output dimensionality compared to the generation of a denoised time domain signal.

## **2.3. Research intent**

The research intent of this project is to investigate the potential of a codebook-based GAN to improve the estimation of noise and speech spectra. Through the results of an extensive search, there has been no direct investigation on the use of GANs to generate optimal codebook inputs in order to enhance low SNR speech signals.

# **3. The proposed approach**

## **3.1.** **Pre-existing works and the proposed approach.**

This study builds on the pre-existing approach from a study [21] conducted by Chehresa et al. To summarise the pre-existing approach, the model is formulated as an overdetermined system consisting of a pre-defined collection of spectral candidates (codebook) and parameters to be estimated based on the current frame of the mixture. Essentially, the following estimation is performed, where the estimated parameters weigh the contribution of each spectral candidate on the final estimate.

From (1-4), the proposed model could potentially result in sub-optimal estimates due to the limited number of spectral candidates, with the possibility of all entries being unable to estimate or capture the spectral profile of the underlying noise or speech spectra, as seen in (1-2) and the resulting least squares solution of the estimated parameters [24]. Therefore, as an extension of this model, this report proposes a new model that results in distinct parameters per frequency bin, as seen in (5), where is the Hadamard product (element-wise multiplication),  is the spectral power for a frequency bin of a codebook entry and is the parameter to be estimated for each frequency bin of a codebook entry. A unique spectral gain function is applied to each codebook spectral candidate.

|  |  |
| --- | --- |
|  | (1) |

|  |  |
| --- | --- |
|  | (2) |

|  |  |
| --- | --- |
|  | (3) |

|  |  |
| --- | --- |
|  | (4) |

|  |  |
| --- | --- |
|  | (5) |

## **3.2. Derivation of codebooks from GMMs**

A Gaussian mixture model (GMM) decomposes a probability density function (PDF) as a multimodal distribution consisting of a mixture of Gaussian distributions. In the context of codebooks, this allows for the compression of all possible noise and speech spectral candidates by clustering similar spectral profiles. To formulate the approximated PDF of the GMM, the following well-known formula is presented where Is the component probability, is the Gaussian operator, and is the component mean and variance, respectively.

|  |  |
| --- | --- |
|  | (6) |

These parameters are derived from the expectation maximum (EM) algorithm, which iteratively re-adjusts the mean, component probability, and variance of each component of the GMM based on the calculated probabilities and initialized parameters until a local maximum-likelihood optimum is reached.

This study models each frequency bin of all data points as a univariate GMM. As the resulting local optimum from the EM algorithm depends on the initial input parameters and no assumptions about the average spectral power per frequency bin were made, sieving was used to derive the codebook entries from the GMM-derived means of each frequency bin. In this study, the number of Gaussian components (, ) are kept the same as the study [21] conducted by Chehresa et al., as seen in (7).

|  |  |
| --- | --- |
|  | (7) |

This study uses sieving to compare GMM models initialized from guessed parameters to acquire the GMM with the highest log-likelihood. The assumption is that the resulting GMM with the highest log-likelihood will be the closest approximation of the true distribution of the sample data or ground truth [25].

As there is the potential of the sample distribution to have a true distribution that cannot be modelled by a mixture of Gaussians (Cauchy or Laplacian distributions), for any non-converging sieved GMMs, the default calculation of the statistical mean is assigned to all the non-converging frequency bins of each spectral candidate in order to avoid number instability from dividing by 0.

## **3.3. Generative adversarial networks**

A generative adversarial network consists of a generator and a discriminator neural network. The discriminator is trained and learns to classify whether the input data is real or fake based on a preset distribution of actual data samples. The generator aims to mimic the distribution of the discriminator by utilizing the discriminator's classification of its outputs to update its parameters per generation/iteration. However, the training stability of this default GAN configuration is known to be very sensitive to the chosen architecture and hyperparameters, which impedes the overall training process [14].

The Wasserstein GAN (WGAN) is a stable derivative of the traditional GAN with the objective of minimizing the Wasserstein or Earth Mover's distance between the data distribution and the generated distribution [26]. As the goal of GANs is to transform a generated distribution () such that is equal to the learned real data distribution of the discriminator (), a metric is needed to measure the difference between , this is known as the Jensen Shannon divergence (JSD), the gradient of which is used to update the parameters of the network to minimize the JSD. However, in situations where does not overlap with , the resulting JSD is a constant value resulting in 0 gradients, which impedes the learning process of the network [27]. Traditional GANs are sensitive to the chosen architecture and hyperparameters, as both will affect and. Compared to the Jensen-Shannon (JS) divergence metric, the Wasserstein distance is an alternative metric that measures the distance of one distribution from another and allows the network to overcome the situation where does not overlap with by minimizing this distance [26]. This results in the reduced sensitivity of the WGAN to the chosen architecture and the chosen hyperparameters.

This study uses a WGAN implementation to generate the required unique parameters; as seen in (5), the proposed model does not have an analytic/model-based solution. Therefore, the generator of the WGAN is configured to be an encoder such that the input power spectral density (PSD) of the mixture is expanded into the exact dimensions as that of the codebook. In this case, the output shape of the generator is NFFT x 9 for noise and NFFT x 6 for speech. The Hadamard product between a GMM-derived codebook and the generated parameters is calculated, and the augmented spectral candidates are summed. The discriminator compares the transformed generated output and adjusts the parameters of the generator. In this study, a semi-supervised approach was taken to pair the input mixture and its resulting generated transform with the ground truth. A high-level diagram of this configuration can be seen in Figure 1.

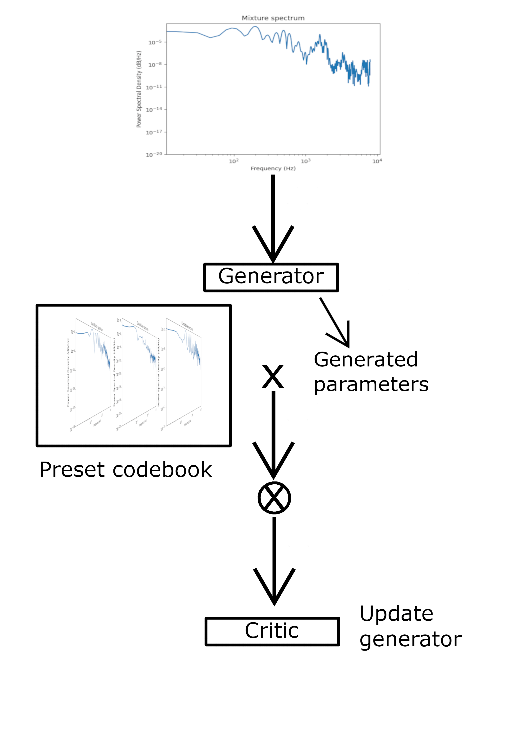


Fig. 1. Proposed WGAN architecture.

The proposed architecture comprises fully connected (dense) layers as the input vector is a PSD of a mixture, and any relationship between each frequency bin value could exist.

# **4. Experiments & results**

## **4.1. Experimental setup**

In order to generate the codebook through a GMM and the proposed GAN, 32ms samples from the VTCK [28] and DEMAND [29] corpora were used. In addition, all WAV audio files were down-sampled to 16Khz through an IIR fifth-order Butterworth low-pass filter. The short-time Fourier transform (STFT) parameters used in this study were 512 samples (32ms) per Hann/Hanning windowed segment, 75% overlap, and 1024-point FFT. The enhanced time-domain signal was resynthesized using overlap-add.

The definition of SNR used in this study is based on the speech presence and the overlapping noise of each mixture determined by a voice activity detector (VAD) [30]. The calculation of the required amplification of noise can be calculated from (8) where is the extracted speech section from a clean file and is the overlapping sections of noise.

|  |  |
| --- | --- |
|  | (8) |

The proposed WGAN algorithm was developed and tested through the TensorFlow platform [31]. In addition, the following hardware was used for development, training, and evaluation: Ryzen 5600X, GTX 1660 Super (6GB) and 16GB DDR4 RAM (3200Mhz).

In order to evaluate the performance of the enhancement algorithm, objective metric scores such as the perceptual evaluation of speech quality (PESQ) [32], short-time objective intelligibility (STOI) [33], and signal-to-distortion ratio (SDR) [34] were used to evaluate the quality, intelligibility and the distortion introduced by the enhancement algorithm.

## **4.2. Results & Discussion**

### *4.2.1. Wiener filtering*

In order to evaluate the noise estimation capability of the proposed WGAN, the performance of the optimal Wiener filter is compared against the semi-optimal Wiener filter. In this evaluation, the PSD of speech is assumed to be known, while the PSD of noise is estimated from the frames of the mixture. From Figures 2-4, the proposed model is shown to nearly match the performance of the optimal wiener filter, given that the PSD of speech is known.

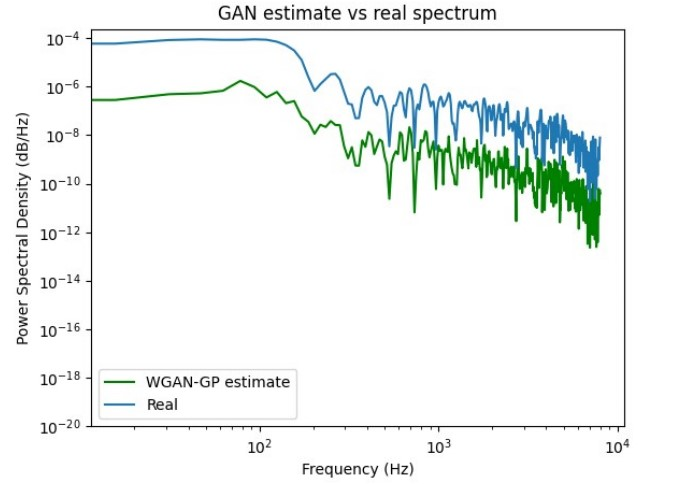


Fig. 2. Estimated spectrum of noise compared against the ground truth.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig. 3. Spectrograms of an unfiltered, WGAN, and optimally Wiener-filtered -9dB waterfall noise audio file.

Table 1: PESQ, STOI, and SDR scores of the semi-optimal Wiener-filtered waterfall audio file.

|  |  |  |  |
| --- | --- | --- | --- |
| SNR(dB) | SDR | PESQ | STOI |
| -3 | -3.747 | 1.047 | 0.552 |
| -6 | -5.602 | 1.042 | 0.532 |
| -9 | -7.611 | 1.032 | 0.512 |

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig. 4. Spectrograms of an unfiltered, WGAN, and optimally Wiener-filtered -9dB babble noise audio file.

Table 2: PESQ, STOI, and SDR scores of the semi-optimal Wiener-filtered babble audio.

|  |  |  |  |
| --- | --- | --- | --- |
| SNR(dB) | SDR | PESQ | STOI |
| -3 | -3.420 | 1.049 | 0.584 |
| -6 | -5.345 | 1.045 | 0.532 |
| -9 | -7.674 | 1.044 | 0.536 |

In order to estimate the PSD of speech as an input to the Wiener filter, the same model was trained to extract the underlying PSD of speech from a mixture. On evaluation, it can be seen from Figure 5 that the proposed WGAN model cannot accurately predict the PSD of speech. Due to this, the filter performs sub-optimally despite the accurate estimation of the noise PSD, as seen in Figure 6, and the resulting objective scores in Table 3 are lower than those presented in Tables 1 and 2.

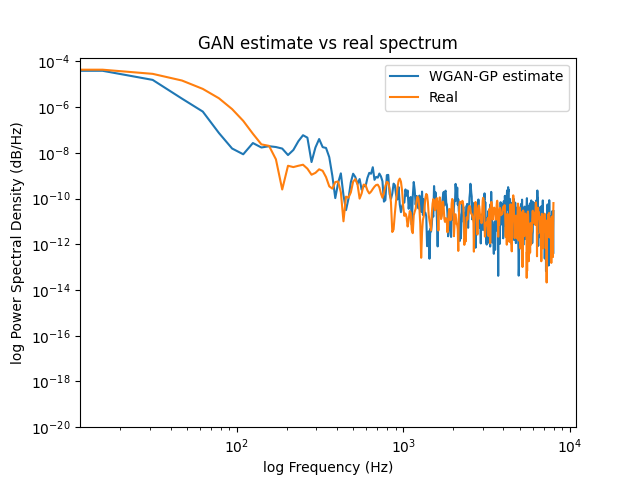


Fig. 5. WGAN estimated PSD of speech against the ground truth.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig. 6. Wiener-filtered babble mixtures with SNRs of -3 dB, -6 dB, and -9 dB (left to right) using the GAN estimated speech and noise PSD.

Table 3: PESQ, STOI, and SDR scores of the Wiener filtered-babble audio file.

|  |  |  |  |
| --- | --- | --- | --- |
| SNR(dB) | SDR | PESQ | STOI |
| -3 | -6.618 | 1.035 | 0.455 |
| -6 | -9.780 | 1.030 | 0.406 |
| -9 | -12.635 | 1.028 | 0.349 |

Spectral power subtraction was also attempted to recover the PSD of speech due to the accuracy of the estimated PSD of noise, but this resulted in over-subtraction due to estimation errors and the minuscule spectral power of speech relative to the spectral power of noise.

As the PSD of speech could not be reliably estimated, the time domain reconstruction of the underlying noise signal from the magnitude spectra was also investigated. This method used the fast Griffin-Lim algorithm (FGLA) [35] to estimate the phase based on the magnitude spectrum of noise and subtract the resulting time-domain signal from the mixture. However, this still yielded suboptimal results and could not reach the performance of the semi-optimal Wiener filter, as seen in Figures 3 and 4.

### *4.2.2. Spectral subtraction*

As an alternative to the Wiener filter, given that only the estimate of noise is available, a well-known spectral subtraction gain function [36, 37] was used, as seen in (9), where is an attenuation factor, is the estimated noise PSD and is the observed mixture PSD.

|  |  |
| --- | --- |
|  | (9) |

From the evaluation of the resulting spectrogram and enhanced audio files, over-attenuating spectral power negatively affected the intelligibility and quality of the enhanced speech signal. Further upward increments of the attenuation factor resulted in the decreased distortion of the enhanced signal, as seen by the decreasing SDR trend from Table 4. However, the over-attenuation of spectral power negatively impacted the intelligibility and speech quality, as reflected by the measured PESQ and STOI scores in Table 4. For higher SNR mixtures, this filter introduced destructive effects on the intelligibility and quality of the resulting enhanced signal from subjective listening sessions.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig. 7. Filtered waterfall noise mixture with an SNR of -9dB and an incrementing alpha value (left to right) using the GAN estimated noise PSD.

Table 4: PESQ, STOI, and SDR scores of the filtered waterfall audio file.

|  |  |  |  |
| --- | --- | --- | --- |
| SNR(dB) | SDR | PESQ | STOI |
| -3 | -12.826 | 1.025 | 0.3707 |
| -6 | -12.810 | 1.025 | 0.3707 |
| -9 | -12.669 | 1.023 | 0.3702 |

### *4.2.3. Reflection on research objectives*

The research objective of this study was to investigate whether GANs could be used to improve the performance of codebook-based Wiener filtering approaches for low SNR speech enhancement through the accurate estimation of the noise spectra. From the semi-optimal Wiener filter results, the proposed approach introduced minimal distortion to the enhanced signal while showing near identical performance to the optimal Wiener filter configuration. However, this configuration is impractical and assumes that the PSD of speech was known at any instant. Attempted alternative methods of speech PSD estimation and filtering showed that the proposed model is not suitable to be used for low SNR speech enhancement due to the failure to accurately estimate the PSD of speech and the resulting introduced distortion to the enhanced signal from alternative filtering methods that utilize only the information provided by the PSD of noise.

Even though the proposed model is shown to fail in estimating the accurate PSD of speech, this study shows that a GAN can predictively generate optimal noise spectra through a codebook to improve further other Wiener filter-based approaches with alternative estimators of speech.

### *4.2.4. Research limitations*

The availability of unseen datasets and hardware limitations limited the findings of this research. Although the proposed model could accurately estimate the PSD of noise, this could be attributed to the lack of other datasets for unseen mixture generation and model evaluation. In order to train and evaluate the proposed model, only mixtures generated from the VCTK and DEMAND databases were used, and potentially different results can be obtained by including other datasets such as TIMIT [38] and NOISEX-92 [39] database. Furthermore, this study could only manage to train on 25% (300000 utterances) of the total non-overlapping frames of the entire dataset due to the limitations imposed by 16GB of RAM.

# **5. Conclusion**

To conclude, although the proposed model could accurately estimate the PSD of noise, it was found that the accurate estimation of the speech PSD also greatly affected the resulting intelligibility of the enhanced signal. Other attempts at filtering and deriving the PSD of speech have shown suboptimal and unintelligible results due to slight errors in the noise estimate, resulting in over-subtraction and distortion.

To summarise, the following contributions have been made:

* In studies concerning Wiener filtering [17, 21], a common assumption is that an accurate noise estimate will result in increased intelligibility and better enhancement performance. This study finds that the sub-optimal or optimal speech estimation also significantly contributes to the intelligibility of the enhanced signal.
* The proposed approach provides insight into the filter characteristics of the proposed WGAN approach through the inspection of all augmented codebook entries and the resulting estimated spectra.

Lastly, potential future directions of this work could consist of further developing the proposed architecture to acquire more accurate noise and speech spectra estimates. In addition, the proposed model could be integrated with other speech estimators to potentially improve the intelligibility of the enhanced low SNR mixtures.

# **References**

[1] Y. Hioka, M. Kingan, G. Schmid and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, 2016, pp. 1-5, doi: 10.1109/IWAENC.2016.7602937.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process., vol. 33, no. 2, pp. 443–445, Apr. 1985, doi: 10.1109/TASSP.1985.1164550.

[3] C. Breithaupt and R. Martin, “MMSE estimation of magnitude-squared DFT coefficients with superGaussian priors,” in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)., Hong Kong, China: IEEE, 2003, p. I-896-I–899. doi: 10.1109/ICASSP.2003.1198926.

[4] S. Gazor and Wei Zhang, “Speech probability distribution,” IEEE Signal Process. Lett., vol. 10, no. 7, pp. 204–207, Jul. 2003, doi: 10.1109/LSP.2003.813679.

[5] A. Kundu, S. Chatterjee, A. Sreenivasa Murthy, and T. V. Sreenivas, “GMM based Bayesian approach to speech enhancement in signal / transform domain,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2008, pp. 4893–4896. doi: 10.1109/ICASSP.2008.4518754.

[6] B. Chen, H. Wang, Y. Wei, and R. H. Y. So, “Truth-to-Estimate Ratio Mask: A Post-Processing Method for Speech Enhancement Direct at Low Signal-to-Noise Ratios,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2020, pp. 7509–7513. doi: 10.1109/ICASSP40776.2020.9052919.

[7] X. Hao et al., “Masking and Inpainting: A Two-Stage Speech Enhancement Approach for Low SNR and Non-Stationary Noise,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2020, pp. 6959–6963. doi: 10.1109/ICASSP40776.2020.9053188.

[8] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks,” IEEEACM Trans. Audio Speech Lang. Process., vol. 26, no. 9, pp. 1570–1584, Sep. 2018, doi: 10.1109/TASLP.2018.2821903.

[9] M. H. Soni, N. Shah, and H. A. Patil, “Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB: IEEE, Apr. 2018, pp. 5039–5043. doi: 10.1109/ICASSP.2018.8462068.

[10] P. C. Loizou and G. Kim, “Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions,” IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 1, pp. 47–56, Jan. 2011, doi: 10.1109/TASL.2010.2045180.

[11] X. Xu, R. Flynn, and M. Russell, “Speech intelligibility and quality: A comparative study of speech enhancement algorithms,” in 2017 28th Irish Signals and Systems Conference (ISSC), Jun. 2017, pp. 1–6. doi: 10.1109/ISSC.2017.7983599.

[12] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, “An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement,” Electronics, vol. 10, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/electronics10010017.

[13] K. Schulze-Forster, “Informed audio source separation with deep learning in limited data settings,” theses.hal.science, Dec. 09, 2021. https://theses.hal.science/tel-03511031/.

[14] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative Adversarial Networks: An Overview,” IEEE Signal Process. Mag., vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[15] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 4396–4405. doi: 10.1109/CVPR.2019.00453.

[16] G. Liu, K. Gong, X. Liang, and Z. Chen, “CP-GAN: Context Pyramid Generative Adversarial Network for Speech Enhancement,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2020, pp. 6624–6628. doi: 10.1109/ICASSP40776.2020.9054060.

[17] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN-2: Studio-Quality Speech Enhancement via Generative Adversarial Networks Conditioned on Acoustic Features,” in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA: IEEE, Oct. 2021, pp. 166–170. doi: 10.1109/WASPAA52581.2021.9632770.

[18] S. Chehresa and M. H. Savoji, “Codebook constrained iterative and Parametric Wiener filter speech enhancement,” in 2009 IEEE International Conference on Signal and Image Processing Applications, Nov. 2009, pp. 548–553. doi: 10.1109/ICSIPA.2009.5478717.

[19] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” IEEE Trans. Audio Speech Lang. Process., vol. 14, no. 1, pp. 163–176, Jan. 2006, doi: 10.1109/TSA.2005.854113.

[20] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments,” IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 2, pp. 441–452, Feb. 2007, doi: 10.1109/TASL.2006.881696.

[21] S. Chehresa and M. H. Savoji, “MMSE speech enhancement based on GMM and solving an over-determined system of equations,” in 2011 IEEE 7th International Symposium on Intelligent Signal Processing, Sep. 2011, pp. 1–5. doi: 10.1109/WISP.2011.6051692.

[22] W. Manamperi, P. N. Samarasinghe, T. D. Abhayapala, and J. Zhang, “GMM Based Multi-Stage Wiener Filtering for Low SNR Speech Enhancement,” in 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Sep. 2022, pp. 1–5. doi: 10.1109/IWAENC53105.2022.9914707.

[23] Y. Yang, C. Bao, and X. Wang, “Codebook-driven speech enhancement using DNN and harmonic emphasis,” in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur: IEEE, Dec. 2017, pp. 149–154. doi: 10.1109/APSIPA.2017.8282018.

[24] S. L. Brunton and J. N. Kutz, “Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control,” Higher Education from Cambridge University Press, May 05, 2022. https://www.cambridge.org/highereducation/books/data-driven-science-and-engineering/6F9A730B7A9A9F43F68CF21A24BEC339#overview (accessed Oct. 12, 2023).

[25] Machine Learning & Simulation, “Sieving & Convergence Analysis for the EM of the Gaussian Mixture Model in Python,”. (Jun 8, 2021). Accessed: Oct. 12, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=Vj4b4xojPMw

[26] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein Generative Adversarial Networks,” in Proceedings of the 34th International Conference on Machine Learning, PMLR, Jul. 2017, pp. 214–223. Accessed: Oct. 11, 2023. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[27] Y.-J. Cao et al., “Recent Advances of Generative Adversarial Networks in Computer Vision,” IEEE Access, vol. 7, pp. 14985–15006, 2019, doi: 10.1109/ACCESS.2018.2886814.

[28] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” Ed.ac.uk, 2019, doi: https://hdl.handle.net/10283/3443.

[29] J. Thiemann, N. Ito, and E. Vincent, “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” Proceedings of Meetings on Acoustics, Jan. 2013, doi: https://doi.org/10.1121/1.4799597.

[30] A. Veysov, “Silero VAD,” GitHub, Mar. 04, 2022. https://github.com/snakers4/silero-vad.

[31] M. Abadi et al., “TensorFlow: A System for Large-Scale Machine Learning This paper is included in the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16). TensorFlow: A system for large-scale machine learning,” Nov. 2016. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[32] 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,” in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752. doi: 10.1109/ICASSP.2001.941023.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech,” IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 7, pp. 2125–2136, Sep. 2011, doi: 10.1109/TASL.2011.2114881.

[34] Cédric Févotte, Rémi Gribonval, Emmanuel Vincent. “BSS\_EVAL Toolbox User Guide – Revision 2.0”. Department of Engineering, Cambridge, United Kingdom, inria-00564760. 2011. Accessed Oct. 12, 2023. [Online]. Available: https://inria.hal.science/inria-00564760

[35] N. Perraudin, P. Balazs and P. L. Søndergaard, "A fast Griffin-Lim algorithm," 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 2013, pp. 1-4, doi: 10.1109/WASPAA.2013.6701851.

[36] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” IEEE Trans. Speech Audio Process., vol. 7, no. 2, pp. 126–137, Mar. 1999, doi: 10.1109/89.748118.

[37] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., vol. 27, no. 2, pp. 113–120, Apr. 1979, doi: 10.1109/TASSP.1979.1163209.

[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT:: acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1,” National Institute of Standards and Technology, Gaithersburg, MD, NIST IR 4930, 1993. doi: 10.6028/NIST.IR.4930.

[39] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, no. 3, pp. 247–251, Jul. 1993, doi: https://doi.org/10.1016/0167-6393(93)90095-3.