university of groningen

campus fryslân

CMGAN-Based Speech Enhancement for Automotive Environments: Targeted Noise Reduction

Ting Zhang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

CMGAN-Based Speech Enhancement for Automotive Environments: Targeted Noise Reduction

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen
Primary supervisor: **Dr. Shekhar Nayak**
External supervisor: **Dr. Nitya Tiwari**
Second reader: **Dr. Matt Coler**
(Voice Technology, University of Groningen)

Ting Zhang (S5690145)

July 9, 2024

Acknowledgements

First, I would like to express my sincere gratitude to my supervisor, Shekhar Nayak, and my external supervisor, Nitya Tiwari, for their significant contributions to my thesis project. Their insightful suggestions and patient guidance during our weekly meetings were truly invaluable. I deeply appreciate their support and encouragement throughout these two months.

I also want to extend my thanks to Phat for his assistance and guidance, both in class and on the thesis project. Additionally, my heartfelt thanks go to all my supportive classmates, who have made this journey enjoyable and rewarding.

One year in Leeuwarden has gone by incredibly fast. Thanks to all the person I met long the way, looking forward to the day our journeys intersect once more.

Abstract

This research employs a Conformer Metric Generative Adversarial Network (CMGAN) model, trained on a tailored in-car noisy speech dataset. The methodology incorporates pre-training, hybrid training, and targeted training to assess the model's performance in speech enhancement tasks. In total, four experiments were conducted to determine the most effective training strategies for the model. Results from these experimental setups confirm that this targeted training approach significantly enhances the ASR system's accuracy and reliability. Particularly when fine-tuned for specific noise conditions, the CMGAN model demonstrated substantial improvements in evaluation metrics, such as the Perceptual Evaluation of Speech Quality (PESQ). Moreover, this study shows that the CMGAN model excels in reducing driving noise but shows less efficacy against street noise and air-conditioner noise. In addition to identifying the most effective training strategies for specific noise datasets, these findings also clarify the relationships between noise types and the effectiveness of speech enhancement. This research concludes that focusing on adaptive and specialized training frameworks can greatly improve ASR performance in real-world noise environments, providing valuable insights for advancing speech recognition technology in practical applications.

Key Words: Speech Enhancement, Generative Adversarial Networks, Noise Reduction, Car Environment

Contents

1 Introduction

For most automatic speech recognition (ASR) systems, maintaining high performance in noisy environments still remains a significant challenge. It is hard to acquire a speech that is free from noise when applying the speech recognition system in real life. Taking the automotive cabin as an example, by applying voice commands on vehicles, driver's reliance on various physical controls while driving can be reduced to great extent. But the ASR system' s performance would be degraded by noise sources within the automotive cabin due to its special environment in which the system is operating. Thus, speech enhancement can be considered as a crucial process tackling the effect of noises on speech.

Many researchers have been working on mitigating the effect of the noise on speech. Due to the complexities of speech signals, It still remains a big challenge to keep speech undistorted while reducing noise and thus limiting the performance of speech recognition systems. Besides, speech enhancement (SE) models often struggle with the diversity and complicity of noise conditions in different application environments. Most existing noise reduction models achieve satisfactory performance on commonly used datasets like LibriSpeech (Panayotov, Chen, Povey, & Khudanpur, 2015) or VCTK (Valentini-Botinhao et al., 2017). These datasets often contain clean, controlled environments that do not fully represent the variability found in real-world scenarios. However, there is a lack of research focused on noise-specific speech enhancement models, which are tailored to address particular types of noise in specific environments. This gap highlights the need for more specialized approaches to effectively enhance speech quality in diverse and challenging noise conditions.

1.1 Research Question and Hypothesis

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> *What kind of speech enhancement techniques or networks can be designed to effectively address the noise robustness problem for Automatic Speech Recognition, thereby improving its accuracy and reliability in noisy environments?*

From which the following subquestions are derived:

- What are the current leading speech enhancement algorithms used in ASR systems?
- Which method is most efficient to train a model under specific noise conditions?
- What types of noise that most negatively impact ASR accuracy in automotive cabin?
- What metrics are most effective for evaluating the performance of speech enhancement methods in noisy conditions?

To answer this question, this study first goes through the existing literature related to advanced speech enhancement models. Among various network architectures, Generative Adversarial Networks (GANs) draw my attention with their capacity to enhance speech by effectively distinguishing and filtering out noise from audio inputs. According to the study of (Donahue, Li, & Prabhavalkar,

2018), operating GANs on log-Mel filterbank spectra instead of waveforms will be more effective in enhancing speech. In the study of (Meng, Li, Gong, & Juang, 2018). It further proposed an adversarial feature-mapping (AFM) method and introduced an additional discriminator network to distinguish the enhanced features from the real clean ones. Besides, Generative Adversarial Networks (GANs) are particularly adept at distinguishing between noise and signal, which allows them to generate clearer speech from noisy environments. This capability is expected to result in substantial improvements in ASR accuracy and reliability, as the enhanced speech produced by GANs will be much closer to clean speech, thereby reducing the errors typically introduced by background noise. Based on this, my hypothesis is that employing Generative Adversarial Networks (GANs) for speech enhancement will significantly improve the noise robustness of Automatic Speech Recognition (ASR) systems.

1.2 Thesis Structure

This thesis is structured to provide a comprehensive exploration of speech enhancement method in aotomotive cabin noisy environments using Generative Adversarial Networks (GANs). Chapter 1 mainly introduces the research topic and outlines the research question and hypothesis. Chapter 2 provides a comprehensive literature review, covering various speech enhancement techniques with a focus on deep learning advancements, particularly GANs. Chapter 3 details the data preparation process, describing the selection and creation of clean and noisy speech datasets in automotive settings. Chapter 4.2 outlines the experimental methodology, explaining the model architecture, the training protocols, and the evaluation metrics used to assess the model's performance. Chapter 5 presents the experimental results, offering quantitative evaluations of the model's performance under different noise conditions and a comparative analysis with baseline models. Chapter 7 discusses the findings in detail, addressing the limitations of the study, and suggesting directions for future research.

2 Literature Review

This chapter provides an overview of relevant literature on speech enhancement. Section 2.1 is about the overview of techniques for improving the quality and intelligibility of speech signals, especially in noisy environments. Section 2.2 discusses signal processing-based methods, such as spectral subtraction and Wiener filtering. Section 2.3 explores deep learning-based methods like Autoencoders and Convolutional Neural Networks (CNNs). Lastly, Section 2.4 further discusses in details about the GAN-based speech enhancement models recently.

2.1 Speech Enhancement Overview

Speech enhancement refers to techniques aimed at improving the quality and intelligibility of speech signals. These techniques are particularly crucial in scenarios where speech signals are degraded by noise, reverberation, babble or other distortions. In real-life scenarios, speech signals are often contaminated by various types of noise, such as inevitable environmental and equipment noises, making speech recognition and comprehension challenging. Therefore, speech enhancement techniques are crucial for improving the performance of speech communication and recognition systems. The primary objective is to enhance the target speech signal while suppressing unwanted noise and interference. The clean speech signal is necessary for applications such as speech or speaker recognition, hearing aids, mobile communication. These techniques can be categorized into four main types: signal processing-based methods, model-based methods, deep learning-based methods, and hybrid methods.

Signal processing-based methods primarily employ techniques like filtering and prediction to reduce noise interference, with classical examples including spectral subtraction (Boll, 1979) and Wiener filtering (Wiener, 1949). Model-based methods use statistical or acoustic models, such as Hidden Markov Models (HMM) (Lee, McLaughlin, & Shirai, 1998) and Gaussian Mixture Models (GMM)(Kundu, Chatterjee, Murthy, & Sreenivas, 2008), to predict clean speech signals by leveraging statistical properties to distinguish between speech and noise. Deep learning methods utilize neural networks to learn the features of speech signals, enabling more effective noise reduction. Common models include Autoencoders (Lu, Tsao, Matsuda, & Hori, 2013a), Generative Adversarial Networks (GANs)(Pascual, Bonafonte, & Serrà, 2017)(Cao, Abdulatif, & Yang, 2022), Recurrent Neural Networks (RNNs)(Valentini-Botinhao, Wang, Takaki, & Yamagishi, 2016), and Convolutional Neural Networks (CNNs)(Park & Lee, 2016). Hybrid approaches combine the strengths of the aforementioned methods to achieve more efficient speech enhancement. These techniques might integrate traditional signal processing methods with deep learning models to leverage the benefits of both approaches.

In this section, apart from these speech enhancement methods, it's also worth mentioning the evaluation metrics that will be applied in the later experiments. For the speech quality evaluation, especially for TTS and SE, the most popular and widely used method is through subjective listening tests. Despite the accuracy and repeatable character of subjective evaluations of SE algorithms, they can be costly and time consuming if conducted under strict conditions, like the inclusion of anchor conditions or large panels of listeners are required. (Recommendation, 2003) Consequently, much effort has been placed to develop alternative objective methods that can reliably predict speech quality and correlate well with subjective assessments. For this reason, researchers have proposed many objective speech quality metrics to estimate the subjective overall speech quality and noise

distortions introduced by representative SE algorithms from various classes.

To evaluate the performance of several objective measures in terms of predicting the quality of noise speech enhanced, (Hu & Loizou, 2007) proposed several objective measures with subjective rating scales: segmental SNR (segSNR)(Hansen & Pellom, 1998), weighted-slope spectral distance (WSS)(Klatt, 1982), Perceptual evaluation of speech quality (PESQ)(Rix, Beerends, Hollier, & Hekstra, 2001), Itakura-Saito distance measure (IS)(Chu & Messerschmitt, 1982), and cepstrum distance measures (CEP)(Kubichek, 1993) and so on. Among these measures, a set of commonly used metrics are chosen to evaluate the enhanced speech quality in the subsequent chapters.

2.2 Signal Processing-based Methods

Before talking about the classic signal processing-based methods, it's worth mentioning that based on the number of acquisition channels involved, the speech enhancement techniques can mainly be divided into two main categories: single channel, dual or multi-channel. The single channel speech enhancement technique uses only one microphone whereas multiple channel speech enhancement technique uses an array of more than one microphone. Single microphone systems use techniques such as spectral subtraction and Wiener filtering, aimed at improving the degraded speech. For many years, spectral subtraction has been taken as the principal way to complete the SE task, in which the power spectrum of an estimate of noise is subtracted from that of noisy speech. Specifically, the input speech is converted into a short-time Fourier transform (STFT) representation, then the noisy amplitude spectrum is modified. In the last step, an inverse STFT is followed by overlap-add synthesis to reconstruct the output signal, as illustrated by the work of (O'Shaughnessy, 2024) Another classical method used for SE, Wiener filtering, uses optimal filtering based on the statistical properties of speech and noise to minimize the mean squared error (MSE) between the estimated clean speech and the actual clean speech.

As mentioned by (Chaudhari & Dhonde, 2015), Multi-microphone systems, on the other hand, use multiple inputs to extract clean speech from noisy and reverberant conditions, utilizing spatial filtering and beamforming techniques. Based on the work of (Priyanka, 2017), Multichannel SE typically assumes that the array of recording microphones is fixed in advance to capture the voices of multiple speakers in a far field. Signals captured by a microphone array can be processed to create a beam pattern through a technique known as beamforming or spatial filtering, for instance, video conferencing or smart homes. In such cases, beamforming is used to amplify the audio that comes from a specific direction or a target speaker. This method enhances target speech by summing weighted and delayed versions of the desired speech signal, while effectively addressing interference audio sources by treating them as independent and incoherent. By utilizing the angle and frequency of incoming signals, beamforming effectively nullifies interference from unwanted directions, hence the term spatial filtering. This technique enhances the amplitude of the speech signal coming from the desired direction.

Based on literature of (Hao, Shan, Xu, Sun, & Xie, 2019), spectral subtraction, Wiener filtering and beamforming, those aforementioned speech enhancement techniques can also be generally classified into statistical-based approaches.

2.3 Deep Learning Methods

Nevertheless, in recent years, neural networks have become the predominant approach for speech enhancement, much like in many other signal-processing applications. This shift is driven by the ability of deep learning to learn complex representations of speech signals, significantly advancing the field. The core idea is to use deep neural networks to capture the characteristics of speech, thereby facilitating noise reduction. Various architectures such as deep autoencoders (DAEs)(Lu, Tsao, Matsuda, & Hori, 2013b), long short-term memory networks (LSTMs)(Weninger et al., 2015), Generative Adversarial Networks (GANs)(Pascual et al., 2017), Recurrent Neural Networks (RNNs)(Valentini-Botinhao et al., 2016), Convolutional Neural Networks (CNNs)(Park & Lee, 2016), and their combinations have been widely adopted, with the incorporation of machine learning strategies like multi-task, progressive, and reinforcement learning.

Deep autoencoders, for instance, are unsupervised learning models that compress input data into low-dimensional representations and then reconstruct the original data, effectively reducing noise interference in speech signals. (Lu et al., 2013b) applied DAE for noise reduction and speech enhancement. The DAE was initially trained using only clean speech, but in subsequent studies, noisy-clean training pairs were introduced to incorporate a denoising process. The final trained DAE, when used as a filter for noisy speech, showed superior performance in noise reduction, speech distortion, and perceptual evaluation of speech quality (PESQ) compared to traditional MMSE-based methods.

In the paper of (O'Shaughnessy, 2024), Shaughnessy further ilustrate the application of CNNs in speech processing. CNNs, typically used for image and video processing, are also widely used in speech enhancement due to their ability to process data within small local regions, which is particularly useful for smoothing variations and reducing data dimensions. In this context, the frequency axis can use Mel-Frequency Cepstral Coefficients (MFCC) instead of simple or reduced STFT, making the rich local patterns present in speech spectrograms.

Similarly, RNNs and LSTMs are designed to capture temporal dependencies within sequence data, modeling the temporal characteristics of speech signals for effective noise reduction. Based on (O'Shaughnessy, 2024), study, except for the short-range time correlations (like vocal tract shapes), SE model also needs to handle information over long-range time correlations, like speaker-specific characteristics that span entire utterances or harmonics over the full spectrum based on fundamental frequency (F0). In the speech enhancement process, the distortion in speech can be present in both short and long-range ways, either abrupt changes or very slowly changing acoustic effects. Thus, RNNs and LSTMs are particularly suited for handling information over these wide data ranges, making them crucial for effective speech processing. (Fu, Wang, Tsao, Lu, & Kawai, 2018) introduced an end-to-end, utterance-based speech enhancement framework employing fully convolutional neural networks. In addition, recent methodologies also applied LSTMs to the denoising task. The study of (Weninger et al., 2015) discriminatively trained LSTM according to an optimal speech reconstruction objective. The study of (Ghosh, Kumar, & Sastry, 2017) incorporated estimated noise characteristics into deep neural network input. The results showed that techniques such as dropout, post-filtering, and the application of perceptual metrics have proven beneficial.

Adding to this, attention-based neural network structures are explored for speech enhancement, inspired by their success in sequence-to-sequence tasks like machine translation, speech recognition, and keyword spotting. The intuitive use of attention mechanisms aims to further enhance speech enhancement performance. In literature, (Hao et al., 2019) proposed an attention-based neural network

approach for single channel speech enhancement. The use of attention mechanisms in SE mirrors human ability to give priority to desired speech signals and thus adjust the focal point dynamically. It is worth mentioning that the attention-based approach can yield a better generalization ability when it comes to unseen noise conditions.

2.4 GAN-based Speech Enhancement

The integration of deep learning into speech enhancement has thus opened new avenues for improving the quality and intelligibility of speech signals, crucial for various applications in modern communication technologies. Another recent breakthrough in the deep learning field are generative adversarial networks (GANs), which have been proven effective in generating realistic images through text-based prompts or by modifying existing images, like converting low-resolution images to high-resolution. GANs consist of a generator and a discriminator, where the generator produces data that mimic real data and the discriminator distinguishes between real and generated data, resulting in clean speech signals. These deep learning models are applied in practical applications such as noise reduction, speech recognition, and speaker recognition. They learn the characteristics of noise within speech signals, separate clean speech from noise, enhance the clarity and intelligibility of speech signals, and extract unique features from each speaker, enabling the identification of different speakers. Thus, this part would go through the main literatures that involve GANs in speech enhancement.

(Pascual et al., 2017) first applied the adversarial framework in speech generation and enhancement tasks. As an end-to-end network working with the raw audio, SEGAN bypasses the extraction of hand-crafted features. Besides, it can learn from multiple speakers and noise types, integrating information into one shared parametrization, enhancing its simplicity and generalizability. Besides, compared to RNNs, SEGAN doesn't require recursive operation, making it process the audio in a quick way. During the GAN training process, D (Discriminator) back-props a batch of real examples, Then, G provides D with a batch of fake examples generated by itself, and D would learn to classify them as fake. Finally, D's parameters are frozen and G back-props to make D misclassify.

Motivated by the success of dealing with image processing tasks in the computer vision field, (Donahue et al., 2018) also evaluated GANs for speech enhancement to boost ASR systems's noise robustness, finding that GANs can effectively enhance speech against additive and reverberant noise. However, Donahue proposed operating GANs on logMel filterbank spectra instead of waveforms. This modification lets GAN require less computation resources and also more robust to reverberant noise. In their experiment, the researchers first trained a SEGAN model in the time domain, and then evaluated the ASR model's performance on noise speech and enhanced speech. The result indicates that SEGAN does not improve ASR performance. They further discovered that operating the SEGAN model on a time-frequency domain improved ASR performance dramatically.

Before taking a look at the new approach of MetricGAN, it's worth mentioning the conditional GANs (cGAN), which is an extension of the traditional GAN. Both the generator and discriminator are conditioned on additional information, which can be anything relevant to the task at hand. In the context of speech enhancement, cGANs can be conditioned on noisy speech to generate clean speech, addressing issues like phase mismatch directly. (Mirza & Osindero, 2014) In particular, for cGANs, except for the adversarial loss, there is an additional LP loss that can guide the learning of generators. Nevertheless, (Fu, Liao, Tsao, & Lin, 2019) found out that the adversarial loss in a conditional GAN (cGAN) is not designed to directly optimize evaluation metrics of a target task, and

thus may not always guide the generator to produce data with improved metric scores. To address this issue, a novel approach called MetricGAN has been proposed. MetricGAN aims to optimize the generator with respect to one or multiple evaluation metrics. Unlike traditional adversarial loss, MetricGAN allows users to specify the desired metric scores for the generated data. By optimizing directly for the evaluation metrics, MetricGAN ensures that the generated speech data achieves higher scores on these metrics, resulting in better overall enhancement performance. In 2021, (Fu et al., 2021) further proposed an improved version of the previous MetricGAN. The MetricGAN+ model applied three training techniques and incorporated domain-knowledge of speech processing. With these techniques, results show that the MetricGAN+ can increase PESQ score by 0.3 on the same VoiceBank-DEMAND dataset compared to the previous model.

In 2022, (Cao et al., 2022) proposed a conformer-based metric generative adversarial network (CMGAN) for SE in the time-frequency (TF) domain. The CMGAN utilizes two-stage conformer blocks in the generator to model both time and frequency dependencies, aggregating all magnitude and complex spectrogram information. Additionally, a metric discriminator is employed to improve the quality of the enhanced speech by optimizing the generator with respect to an evaluation score. Quantitative analysis on the Voice Bank+DEMAND dataset indicates that CMGAN outperforms various previous models, achieving a PESQ (perceptual evaluation of speech quality) of 3.41 and an SSNR (segmental signal-to-noise ratio) of 11.10 dB.

Based on the aforementioned studies, this research focuses on developing and evaluating a Conformer Metric Generative Adversarial Network (CMGAN) model for speech enhancement in specific noisy environments. Using a customized in-car noisy speech dataset and implementing focused training strategies (pre-training, hybrid training, and targeted training) to improve the model's performance in multiple noise conditions.

3 Data Preparation

One of the most significant applications of speech enhancement techniques is the preprocessing of Automatic Speech Recognition (ASR) systems, which play a crucial role in the development of smart cabins. For instance, drivers can give a speech command to adjust the air conditioner temperature to its voice assistant without getting distracted from manual operation while driving. However, the in-car noise environment is much more complex than typical noisy speech datasets. Various noise conditions must be considered, such as the position of the microphones, road conditions (city highway or expressway), different speech speeds, multiple weather conditions, different car types, and other factors.

To address these complexities, a customized dataset for in-car speech enhancement tasks is proposed. Collecting speech data in an actual in-car noisy environment is typically costly; thus, it is a common practice to create artificial noisy speech by adding noise to clean speech recordings. Although this method does not perfectly replicate the complex real road conditions, it provides a practical alternative for conducting this research. Nevertheless, artificially generated signals cannot fully simulate the intricate environment inside the car due to numerous variables that significantly influence noise levels.

To tackle this issue, three datasets will be utilized to generate the in-car speech dataset: VCTK speech dataset(Valentini-Botinhao, Wang, Takaki, & Yamagishi, 2016), Urbansound dataset(Salamon, Jacoby, & Bello, 2014), and noise data in in-car scenes(Hou, Liu, Zhang, & Huang, 2011). The Urbansound dataset contains urban noise from ten different urban noise condition classes, of which four noise types related to in-car environments are selected. The noise data in the in-car scene provides real noise recordings on the road, with specific descriptions of car model specifications and weather conditions. These five types of noise will be used to enrich clean speech with real noise for purposes such as model training and noise characteristic analysis. The VCTK dataset, an open-source speaker-dependent speech dataset, will serve as the foundation for clean speech.

3.1 Data Description

3.1.1 Clean Speech Dataset - VCTK

VCTK speech dataset (Valentini-Botinhao et al., 2016) is provided and published by University of Edinburgh. School of Informatics. Center for Speech Technology Research (CSTR). The database, which is especially designed to train and test speech enhancement methods, consisted of multiple speakers and diverse noise conditions. They provided a clean and noisy parallel speech database.

In my case, I selected the 28 speakers clean subset for making the training set and 2 speakers clean subset for making the testing set. The clean speech dataset was selected from the Voice Bank corpus, which contains 28 speakers: 14 male and 14 female of the same accent region (England). There are around 400 sentences available from each speaker. All data is sampled at 48 kHz and orthographic transcription is also available. However, during the data preparation process, I converted the sampling rate into 16kHz to guarantee the consistency of sampling rate.

3.1.2 Noisy Speech Dataset - Urban Sound

To create the noisy database used for training and testing, I used the in-car noise real recording database as well as the UrbanSound database. Both datasets are open-source and available.

The UrbanSound dataset (Salamon et al., 2014) is a comprehensive collection designed specifically for urban sound research, developed by researchers at the Music and Audio Research Laboratory and the Center for Urban Science and Progress at New York University. UrbanSound encompasses 27 hours of audio recordings, with 18.5 hours of these being annotated to mark occurrences of various urban sound events. This dataset is available for download and is intended for use in research to advance the field of audio analysis, particularly in urban environments. These resources provide a valuable foundation for exploring and modeling urban acoustic phenomena.

This dataset comprises 8732 labeled sound excerpts, each no longer than 4 seconds, derived from 10 urban sound categories: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. These classes were chosen based on their prevalence in urban noise complaints, except for 'children playing' and 'gun shot', which were added to diversify the dataset. This rich collection serves as an essential resource for developing and testing sound classification systems aimed at understanding urban soundscapes. During these 10 classes of urban noises, this research only selected 4 mostly common types of noise while driving: the air conditioner, car horn, children playing and engine idling.

3.1.3 Noisy Speech Dataset - In-car Noise Database

The dataset (Hou et al., 2011) comprises over 500 hours of in-car noise data collected to support detailed noise modeling for automotive applications. It takes into consideration an array of factors such as different vehicle models, road types, vehicle speeds, and window positions in various conditions, making it a valuable resource for voice tech related research and speech enhancement tasks particularly in vehicle cabin settings.

Recordings were made using a combination of four microphones and two mobile phones, resulting in high-quality audio files (microphones at 32 kHz, 32-bit, mono; mobile phones at 16 kHz, 16-bit, mono). These devices were placed at six key points within the vehicle to ensure a thorough sampling of its acoustic environment.

The data covers various weather conditions, including 317 hours on sunny days and 214 hours on rainy days. The documentation also provides us with the detailed annotations that describe the road type, weather conditions, air conditioning status, and other relevant information which might influence the in-car noise level. This dataset is categorized into five main conditions, each containing multiple scenarios tailored for diverse research uses, for instance, developing noise reduction models to improve ASR system's performance within vehicles. This dataset is accessible for academic research, proving a comprehensive tool for developing automotive AI related product development.

3.2 Data Preprocessing

To make the noisy training set, a total of 40 different conditions are considered (ten noises x four SNRs): 5 types of noise (2 noises are selected of each noise type) with 4 signal-to-noise ratio (SNR). The SNR values used for training were: 15 dB, 10 dB, 5 dB and 0 dB. And the noises are combined with the clean speech based on randomly chosen SNR values from these four values.

Specifically, the in-car noises are mostly composed of the following five aspects.

1. **The driving noise**
   The noises while driving are affected by the real driving condition, which is related to the weather, road condition, speed, model type and so on.

2. T**he car horn**
   The honking noise of cars is one of the most common noises in the traffic.

3. **The engine**
   The noise generated by the engine is usually louder if the car is accelerating or traveling at a higher speed than keeping a fixed lower speed.

4. **The air conditioner**
   If the air conditioner is too close to the microphone, the wind from the air conditioner will certainly be caught by the recording devices.

5. **The street noise**
   Including rain drops on the windshield or other vehicles or pedestrians passing by. These kinds of noises will be able to get into the car if we do not close the windows tightly.

To create the noisy database used for testing, I also selected 10 noise conditions from the aforementioned noise types but with slightly higher SNR values: 17.5 dB, 12.5 dB, 7.5 dB and 2.5dB. This created 40 different noisy conditions, but still made sure that the speaker of the clean speech as well as the noisy conditions are completely unseen by the trained model . The noise was added following the same procedure described previously. All the clean speech and noise audio are converted into 16 bit, 16000 Hz, mono channel audio files.

4 Methodology

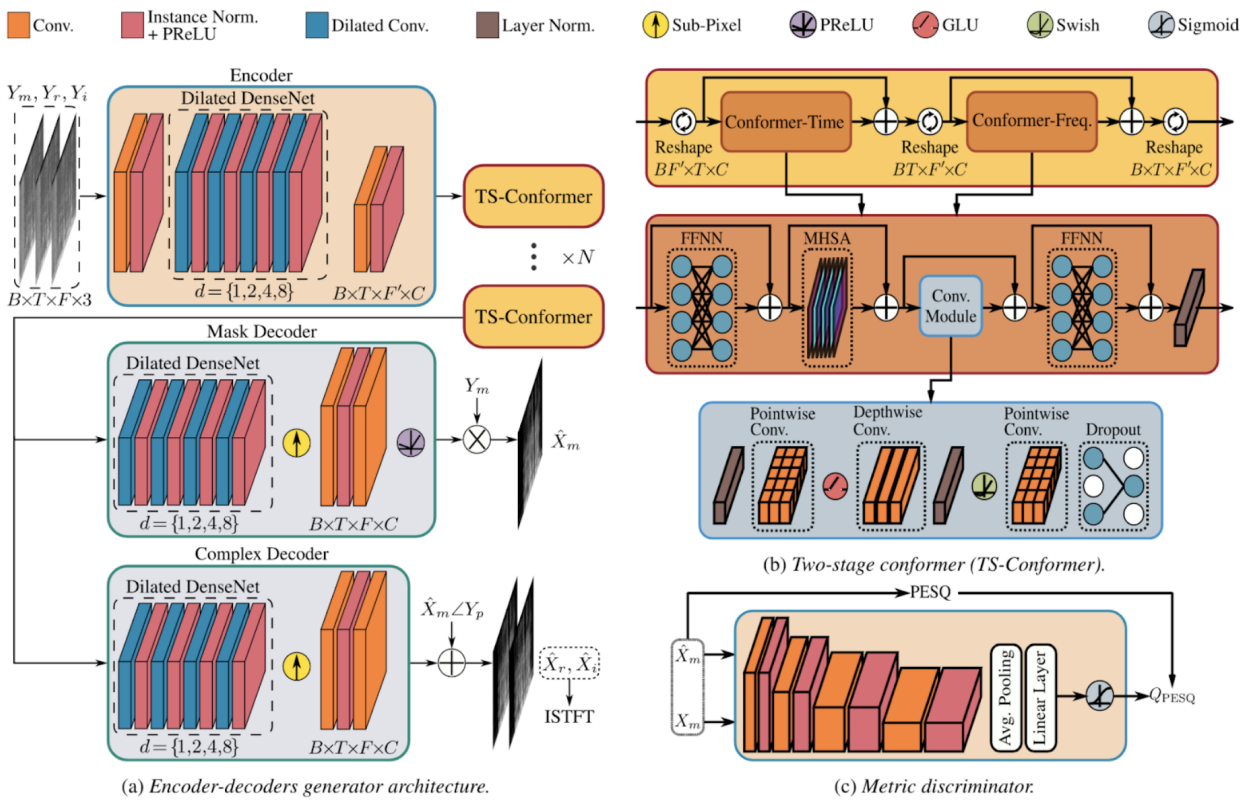
This chapter will provides an overview of the methodology part of the experiment and can be mainly described as follows: Section 4.1 first explains the architecture of the convolution-augmented transformer (Conformer) based metric generative adversarial network (CMGAN) for speech enhancement in the time-frequency (TF) domain. Section 4.2 will describe the experiment setup protocol used for the model training. Besides, section 4.3 will further discuss how to evaluate the fine-tuned model performance on the customized in-car noise speech dataset.

4.1 Model Architecture

Inspired by the work of (Cao et al., 2022), this study replicated and fine-tuned the CMGAN's model architecture and applied it in the in-car noisy reduction scenario. The CMGAN framework basically can be divided into three parts: an encoder-decoder generator and a discriminator.

The encoders use a Dilated DenseNet with dilation factors of d=1,2,4,8, which helps capture contextual information at different scales. This architecture processes concatenated magnitude and complex inputs through the generator, which features an encoder with two-stage conformer blocks. The encoder's role is to create a compact representation of the input features. The mask decoder is tasked with determining the mask for the input's magnitude, while the complex decoder adjusts for the real and imaginary components of the signal. In the meanwhile, the discriminator evaluates a non-differentiable, black-box metric. The aim of the discriminator in CMGAN is to mimic the metric score and take it as a part of the loss function. Before we go further into the details of the model architecture, it's worth mentioning why we tend to apply the MetricGAN in the time-frequency (TF) domain in this case. Recently more studies have been using time domain methods in the SE, in other words, using raw waveforms directly to estimate clean waveforms from the noisy speech. However, in this case, there might be artifacts in the reconstructed speech, since the whole information is processed in the format of waveform. The absence of frequency representation could influence the way models capture speech phonetics in the frequency domain. (Wang, He, & Zhu, 2021)

Unlike the time-domain methods which take the raw waveform as input directly, the TF methods transform raw waveforms into a TF-based representation (e.g. spectrograms) by applying a Short-Time Fourier Transform (STFT) to the inputs. This conversion first breaks the signal into overlapping windows and analyzes each for its frequency content. The resulting complex-valued spectrum provides both magnitude and phase information for each segment. The magnitude reflects the strength of frequency components at specific time points, while phase describes the sine/cosine phase of each frequency, related to their positional attributes in the cycle of the waveform. This dual representation is inclusive and more descriptive compared to raw waveforms, allowing more intricate details of the audio signal to be captured and analyzed.



(a) *Encoder-decoders generator architecture.* (b) *Two-stage conformer (TS-Conformer).* (c) *Metric discriminator.*

Figure 1: Model Architecture (Cao et al., 2022)

4.1.1 Generator

As shown in the figure1, the generator architecture comprises 3 main parts: Encoder, Decoder (Mask decoder and complex decoder), and conformer block.

The generator begins by taking a noisy speech waveform and converting it into a complex spectrogram using the Short-Time Fourier Transform (STFT). The processed spectrogram is split into magnitude, phase, real and imaginary components, which are then combined and used as input features for the encoder. This transformation changes the audio from time-domain into frequency-domain, making it easier to manipulate different frequencies present in the audio. Once in the frequency domain, the spectrogram undergoes power-law compression (C. Kim & Stern, 2009), which adjusts the magnitude of different frequencies based on a compression exponent:

$$Y = |Y_o|^c e^{jY_p} = Y_m e^{jY_p} = Y_r + jY_i \tag{1}$$

This formula describes how to convert the raw speech waveform to a complex spectrogram Yo (a complex matrix with time and frequency dimensions) by STFT, and then further get the compressed spectrogram Y through the power-law compression. The basic principle of power-law compression is to transform each component of the spectrum (usually amplitude) according to a power-law relationship. As we can see in the formula, c is the compression exponent, its value usually between 0 and 1, in our model, the value of this exponent c is set as 0.3, according to Braun et al. This value can be used to adjust the weight of large values while increasing the effect of small values. The

reason behind this power-law compression is that, in the human auditory system, our perception of sound at different volumes is non-linear. When hearing low volume sounds (such as whispers or distant sounds), to which our ears and brain are usually very sensitive. Our perception system would automatically amplify the sound volume and make it easier to be heard. On the contrary, the high volume sounds will not increase the perceptive intensity in the same proportion. In terms of speech enhancement experiments, this step is crucial for enhancing quieter sounds, making the dynamic range of the entire signal more in line with the characteristics of human hearing.

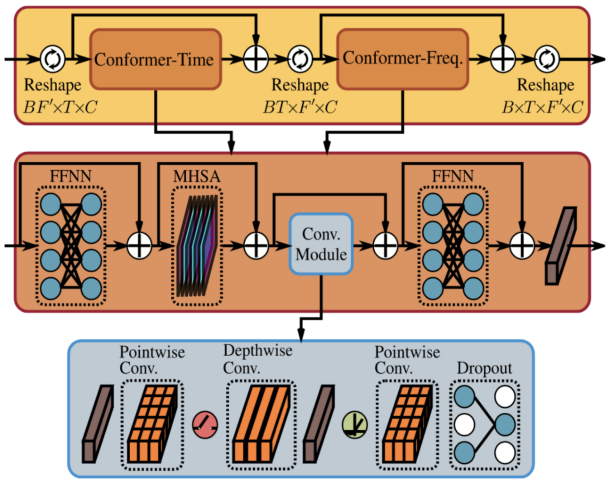
Encoder The encoder consists of two convolution blocks and a dilated DenseNet structure. The convolution blocks are responsible for extracting more complex features of input features, enhancing different aspects of the audio signal by capturing both local features (using convolutions) and broader features (using densely connected blocks). As for the dilated DenseNet, it can efficiently expand the receptive field with the dilation factors of 1,2,4,8. This allows the encoder to capture a wider range of context information without significantly increasing the number of parameters.

Decoder After encoding, the architecture splits into two paths: Mask Decoder and Complex Decoder. The mask decoder uses another Dilated DenseNet to generate a mask that will be applied to the magnitude. Similar to the Mask Decoder, the complex decoder operates to generate a mask for the complex components (real and imaginary parts). This helps in reconstructing the original speech signal with enhanced quality. The outputs from these decoders are combined with the original inputs to produce a refined spectrogram, which is then transformed back to the time domain using the Inverse Short-Time Fourier Transform (ISTFT).

Two-stage conformer block In recent years, transformers, which can effectively handle long-term dependencies using multi-head self-attention (MHSA), have demonstrated superior performance in sequence-to-sequence fields. They have been particularly successful in speech-related applications such as ASR, SE, sound event detection and speech separation for their capability of capturing long-distance dependencies. Meanwhile, CNNs can exploit local features effectively. As the combination of CNNs and transformers, more recently, convolution-augmented transformers (conformers) have gained people's attention with their enhanced ability to reflect the local and global temporal context dependencies. (E. Kim & Seo, 2021)

In the architecture of CMGAN, it employs two conformer blocks sequentially in order to process the encoded features in both time and frequency dimensions. Generally speaking, the first conformer block focuses on temporal dependency. After temporal features are processed, the second block block will focus on the frequency-based features instead, allowing for a comprehensive understanding of the speech signal. The process starts with the time and frequency reshaping. The feature map is reshaped to emphasize either time or frequency aspects before entering each conformer stage. This step is crucial because it prepares the data in a form that is more suitable for the specific focus of each conformer block as we mentioned earlier. By doing this, the model can more effectively concentrate on one aspect of the data at a time.

The conformer architecture is shown in the figure2. For each conformer block, it includes two feed-forward neural networks (FFNN), between these two FFNNs is a multi-head self-attention (MHSA) featuring 4 heads, which enhances the model's ability to focus on relevant features across different positions in the input sequence. Following this attention mechanism is a convolution module. As we can see in the figure2, this module begins with a layer normalization. Following this,



(b) *Two-stage conformer (TS-Conformer).*

Figure 2: Conformer Architecture (Cao et al., 2022)

a point-wise convolution layer takes over, applying a convolution operation to each point independently across the depth, effectively mixing channels without considering spatial or temporal details, which is crucial for fine-tuning channel-specific features. Subsequently, a Gated Linear Unit (GLU) Activation introduces non-linearity which can help with the common vanishing gradient problem in DNNs. The sequence continues with a 1D-Depthwise Convolution Layer, where each input channel is independently processed by separate kernels. In this case, this layer uses a Swish activation function. It combines the advantages of ReLU and sigmoid functions to facilitate smoother gradient flows during backpropagation. Another Point-wise Convolution Layer follows, further refining the feature representation by mixing the channels once more. The module concludes with a Dropout Layer. This comprehensive sequence within the convolution module ensures robust feature processing, which can improve the model's overall performance. Moreover, a residual connection is employed, linking the input directly to the output of the conformer block. (Cao et al., 2022)

4.1.2 Metric Discriminator

In the model training process, to assess the effect of improvement of sound quality, specific assessment metrics (such as PESQ) would usually be applied in model evaluation. However, here comes the problem, some evaluation metrics often cannot be used directly as "loss functions" in the optimization process to adjust the model, because they are mathematically non-differentiable, for instance, perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI). This implies that at certain points, or potentially across its entire domain, a function lacks a clear derivative. When a function is non-differentiable at a particular point, it means a gradient cannot be established there. Consequently, we are unable to employ gradient descent or similar optimization methods to directly enhance the function, in other words, making it difficult for the model to optimize these metrics directly.

Therefore, the primary role of a metric discriminator is to approximate and optimize for metrics that are otherwise non-differentiable and cannot be directly incorporated into the loss function of a

model. The metric discriminator learns to predict the value of these complex metrics from the model outputs. By doing so, it effectively translates these metrics into a form that can be used in the model optimization. Once the metric discriminator has learned to predict the metric reliably, its predictions are used as part of the overall loss function. This allows the primary model (like a generator in a GAN) to be trained not just to minimize a traditional loss, but to optimize towards improving the actual perceptual quality of the output, as measured by the metric.

In the context of CMGAN, the metric discriminator is specifically tasked with approximating the PESQ score. It evaluates both ”clean” (original) and ”enhanced” (processed by the generator) audio samples to determine how close the enhanced version is to the original in terms of perceived quality. The feedback from the metric discriminator is then used to adjust the generator, pushing it towards producing enhancements that get higher PESQ scores.

As for the specific architecture of the discriminator, it is composed of four convolutional blocks. Each block begins with a convolutional layer, which is succeeded by instance normalization and a PReLU activation function. Following these convolution blocks, global average pooling is performed, leading to two feed-forward layers capped with a sigmoid activation function. (Cao et al., 2022)

4.2 Experiment Setup

The foundational concept of this experiment derives from the intention of improving ASR performance in automotive cabin noise environments by applying speech enhancement techniques. In line with this intention, here comes the thought of whether integrating specific in-car noise into clean speech data to form a new dataset would enable the CMGAN model to assimilate the characteristics of these noise types and thereby improve speech enhancement (SE) performance.

Our hypothesis is that fine-tuning the CMGAN model on a custom VCTK dataset, built upon a pre-trained model, could yield superior performance across various evaluation metrics.

This hypothesis is predicated on the assumption that tailored model training under specific noise conditions can more effectively capture and mitigate the impact of such noise conditions on speech quality. Consequently, the subsequent section is dedicated to detailing the experimental setup and validating the efficacy of this custom model training approach. The experimental framework is structured into the following sections: pre-trained model on VCTK Demand dataset, model trained from scratch on customized dataset, hybrid training, comparative analysis based on noise conditions.

4.2.1 Pre-training

The experiment starts with training the model on the standard VCTK demand dataset as the pre-trained model. This model has already been trained on a larger set of voice data, allowing it to have a broad understanding of speech characteristics in various acoustic conditions. Using this pre-trained model as a baseline can help us establish a reference point for evaluating the enhancements brought by subsequent training phases. It allows us to measure the effectiveness of our customized modifications against a known standard in speech processing, giving us a clear benchmark for comparison.

4.2.2 Targeted Training

The second phase of our experiment involves developing a new model from scratch, trained exclusively on our specially created in-car noisy speech dataset. This approach is crucial because it lets us see how a model, built without prior knowledge or biases from general voice data, learns and adapts to specific noisy environments typical in automotive settings. Training a model from scratch on this dataset enables us to explore the unique characteristics and challenges presented by in-car noise, such as varying levels of background sounds and different types of disturbances that can affect speech clarity. By comparing this model's performance to our baseline, we can estimate the benefits or limitations of training solely on targeted noise conditions without the influence of pre-existing model parameters.

4.2.3 Hybrid Training

In the hybrid training phase, the model goes through a two-stage training process. First, it's trained on the VCTK-demand dataset for 75 epochs, using a wide range of speech data to build a strong basic understanding of speech patterns. This stage aims to create a robust model that can handle different acoustic environments generally. After this pre-training, the model is fine-tuned on a customized noisy speech dataset from epoch 76 to epoch 119. The fine-tuning specifically targets noise conditions like those we encountered while driving in cars, which are crucial for practical application of the ASR system like voice commands. This approach helps the model learn not just general speech patterns but also to recognize and handle specific noise types, improving its performance in multiple noisy settings.

4.2.4 Comparative Analysis Based on Noise Conditions

The comparative analysis tests how well the model works across various real-world conditions. After choosing the best model from the initial setups, this phase evaluates its performance across 20 different noise subsets. Each subset is created by adding different types of car noises—such as driving noise, car horns, engine sounds, air conditioners, and street noise—into clean speech samples at different noise levels: 15 dB, 10 dB, 5 dB, and 0 dB. To set up these conditions, 100 audio files of clean speech from the same speaker are selected from the VCTK dataset. These files are then mixed with noise at specified levels to simulate challenging acoustic environments. This process ensures each mixed audio file in a subset represents a potential real-life scenario. Each subset is then systematically named (e.g., subset_1_type1_snr15) to keep the testing organized.

This structured approach allows us to measure the model's ability to improve speech under various noise types and see how different noise characteristics and levels affect speech quality. By examining the results across these conditions, we can pinpoint specific areas where the model excels or falls short, guiding further improvements and the development of more effective speech enhancement methods. This comprehensive experimental design will not only help in identifying the most effective model configuration but also enhance our understanding of the relations between different in-car noise conditions and their impact on speech enhancement outcomes. Which hopefully could contribute to the development of more adaptive and efficient noise-robust speech enhancement technologies.

Evaluation Metrics	Explanation	Scale	Evaluation target
PESQ	Perceptual Evaluation of Speech Quality	-0.5 to 4.5	Perceptual Quality
CSIG	Mean opinion score of signal distortion	1 to 5	Perceptual Quality
CBAK	Mean opinion score of background noise	1 to 5	Perceptual Quality
COVL	Mean opinion score of overall effect	1 to 5	Perceptual Quality
SSNR	Segmental Signal-to-Noise Ratio	-	Intelligibility
STOI	Short-Time Objective Intelligibility	0 to 1	Intelligibility

Figure 3: Evaluation Metrics

4.3 Evaluation Metrics

To evaluate the performance of several objective measures in terms of predicting the quality of noise speech enhanced, (Hu & Loizou, 2007) proposed several objective measures with subjective rating scales, of which a set of commonly used metrics are chosen to evaluate the enhanced speech quality. As shown in the table, these metrics include the Perceptual Evaluation of Speech Quality (PESQ), which ranges from -0.5 to 4.5, and the Segmental Signal-to-Noise Ratio (SSNR), which for practical purposes spans from -10 to 35dB, focusing on the power ratio between speech and noise within frames where speech is detected. In the experiment I also use Mean Opinion Score (MOS)-based metrics: MOS prediction of the signal distortion (CSIG), MOS prediction of the intrusiveness of background noise (CBAK), and MOS prediction of the overall effect (COVL), all scored from 1 to 5. Additionally, the Short-Time Objective Intelligibility (STOI) metric, which ranges from 0 to 1, is used to evaluate speech intelligibility by comparing overlapping frames of clean and enhanced speech signals, applying normalization and clipping to mitigate the impact of outliers. It's worth noting that for all the metrics, the higher value the better performance.

5 Results

This chapter presents and interprets the results of the experiments described earlier, following the structure of the experiment setup section. Section 5.1 evaluates the models trained on the original VCTK-demand dataset and their performance on both the VCTK-demand and a customized in-car speech dataset through comparative analysis. Section 5.2 discusses the training processes of models developed from scratch for the in-car speech dataset, including the selection of the best checkpoint for further evaluation. Section 5.3 validates the hypothesis of performance enhancement in the fine-tuned model. Finally, Section 5.4 analyzes the impact of various noise conditions introduced in the customized in-car speech dataset, aiming to identify which noises most significantly affect the speech enhancement algorithm's performance.

5.1 Experiment 1 Result: Pre-trained Model

In the experiment 1, the pre-trained model on the VCTK Demand dataset serves as a baseline for subsequent evaluations. This experiment assessed the performance of models trained for 25, 50, and 75 epochs on both the VCTK-demand and in-car noise datasets, revealing how model performance adapts to different acoustic environments. As shown in the accompanying table, performance metrics typically get worse when the model is applied to the in-car noise dataset compared to the VCTK-demand dataset.

Specifically, As illustrated in the table4, the PESQ scores, which assess perceptual speech quality, show a consistent decline of more than 0.1 when transitioning from the VCTK-demand to the in-car noise dataset. This significant drop highlights the challenges posed by the in-car noise conditions. Similarly, the metrics for CBAK, COVL, SSNR, and STOI also show slight declines, indicating a general decrease in both noise suppression effectiveness and overall speech intelligibility under more challenging noise conditions.

As for the CSIG scores, which measure signal distortion, increase slightly in the in-car noise dataset. This could suggest that while overall noise characteristics are more complicated in the in-car environment, the specific type of noise may allow the speech enhancement algorithm to maintain or even slightly improve signal clarity relative to the background noise.

Overall, these results underscore the impact of noise conditions on speech enhancement algorithms. The outcome of experiment 1 demonstrates that applying the model, which was pre-trained on the larger VCTK-demand dataset, to evaluate the customized in-car speech noise dataset did not yield satisfactory results. This further highlights the necessity of the subsequent experiments, which will be discussed in the following section.

Pre-trained model	Training set	Evaluation set	PESQ	CSIG	CBAK	COVL	SSNR	STOI
25 epochs	VCTK-demand	VCTK-demand	1.68	2.59	2.43	2.16	1.46	0.85
25 epochs	VCTK-demand	In-car noise dataset	1.56	2.67	2.36	2.13	1.47	0.84
50 epochs	VCTK-demand	VCTK-demand	1.53	2.41	2.36	1.99	1.45	0.82
50 epochs	VCTK-demand	In-car noise dataset	1.43	2.45	2.30	1.96	1.39	0.81
75 epochs	VCTK-demand	VCTK-demand	1.58	2.43	2.38	2.03	1.44	0.84
75 epochs	VCTK-demand	In-car noise dataset	1.48	2.56	2.32	2.04	1.43	0.82

Figure 4: Pre-trained Model's Performance

5.2 Experiment 2 Result: Model Trained on Customized Dataset

In the second experiment, the objective was to assess the performance of training a model separately on the new in-car noise dataset without using any pre-training technique on the larger dataset with general noise types like VCTK-demand dataset. This approach aimed to determine if a model, when trained exclusively on domain-specific data, could still get competitive performance.

The experiment involved training the model from scratch, closely monitoring the generator and discriminator losses throughout the process, as illustrated in the figure6. Over a duration of 72 hours, the model trained 75 training epochs, demonstrating a promising trend in the stabilization of loss values. These losses initially fluctuated but gradually became stable. Particularly notable was the performance after 21 epochs, where the model reached its optimal effectiveness across all measured metrics. The performance of this model was quantitatively superior, with the following scores:

PESQ=2.33, CSIG=3.83, CBAK=2.85, COVL=3.12, SSNR=2.57, STOI=0.91

However, in the later epochs, the model's performance declined, as indicated by the evaluation metrics displayed in the figure 6. Given the superior performance of this model compared to others trained for different durations, it was selected as the best checkpoint for further detailed analysis. In Experiment 4, this model will be used to conduct a comparative analysis across multiple noise conditions, aiming to validate its generalizability and effectiveness in real-world scenarios.

Except for the evaluation metrics, the model with best performance (epoch-21) is used to conduct a spectrogram comparative analysis of enhanced speech effect. As shown in the figure 7, the spectrogram analysis provided offers a visual comparison of a speech utterance before and after processing by the speech enhancement model. The clean spectrograms display clear harmonic structures indicative of the fundamental frequency and its harmonics, free from noise disruptions. In contrast, the noisy spectrograms exhibit significant smearing and speckling across all frequencies, obscuring the clarity of speech components and highlighting the disruptive effect of background noise. The enhanced spectrograms show a substantial reduction in noise, with the fundamental and harmonic

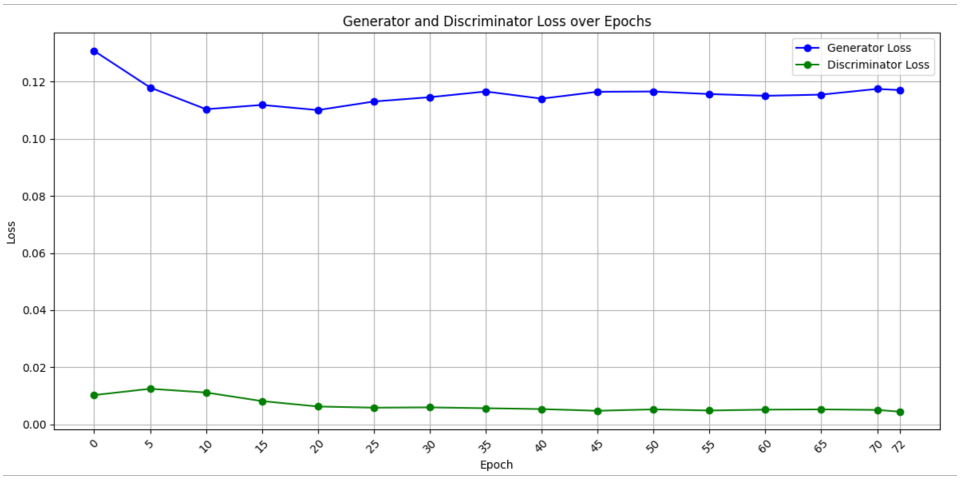


Figure 5: Generator Loss & Discrinator Loss over Epochs

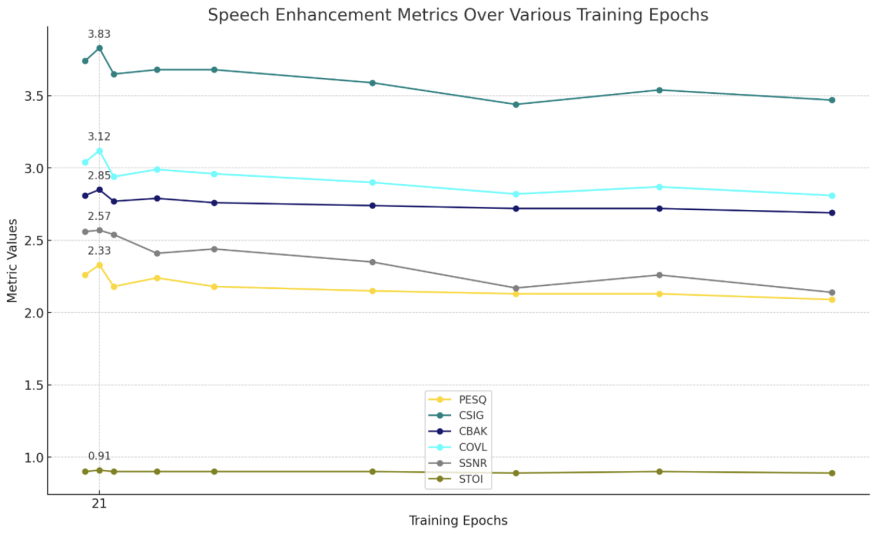


Figure 6: Speech Enhancement Metrics over Epochs

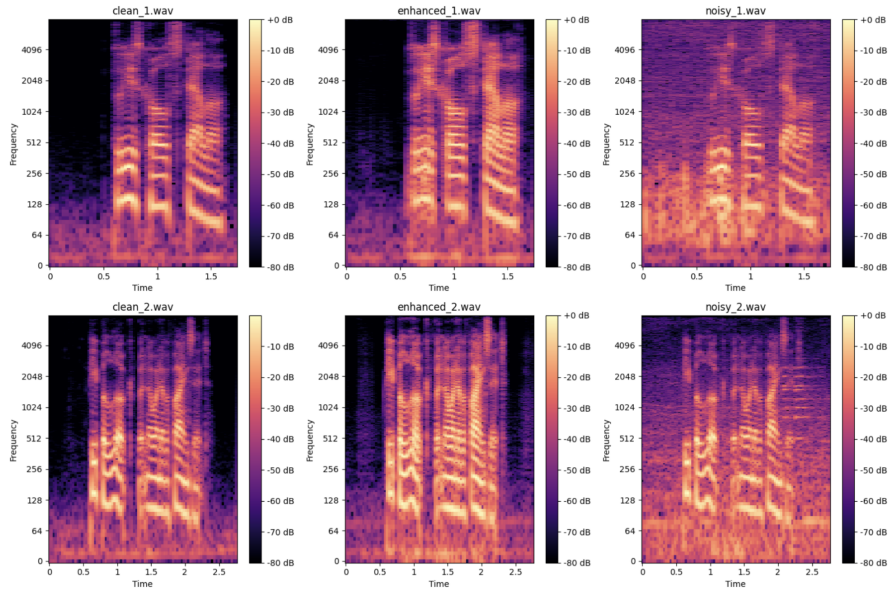


Figure 7: Spectrograms of utterances in the best track that saved: clean target, enhanced speech, noisy input

structures more visible, though some residual noise is still present, especially in higher frequencies. This demonstrates the model's effectiveness in mitigating noise and improving the clarity of speech signals, while also pointing to areas for further refinement to achieve even clearer and more intelligible speech output.

5.3 Experiment 3 Result: Fine-tuned Model

The experiment 3 conducted a hybrid training approach that involved a two-stage training process. Initially, the model that underwent 75 epochs of training in the first experiment served as the baseline. This baseline model was originally trained only on the VCTK-demand dataset. From epoch 76 onward, up to epoch 119, this model was fine-tuned on a customized noisy speech dataset specifically designed to simulate automotive noise conditions. The objective of this experiment was to assess the effectiveness of fine-tuning on a model's performance in specific noise environments.

The evaluation metrics for the baseline model were as follows:

PESQ=1.48, CSIG=2.56, CBAK=2.32, COVL=2.04, SSNR=1.43, STOI=0.82

These values of the baseline model serve as the benchmark for observing potential improvements after the fine-tuning process. According to the results displayed in the figure 8and table9, there was a marked improvement in all metrics during the initial 10 epochs of fine-tuning (epochs 76 to 85). The CSIG score, for instance, soared from 2.56 to 3.50, indicating a substantial enhancement in signal quality. Conversely, the STOI score showed minimal improvement, suggesting that the intelligibility of speech, though slightly better, was not as dramatically affected by the fine-tuning as other metrics.

Following the initial surge, the metrics continued to improve or stabilize at high levels, with the peak performances occurring around epoch 85. After that stage the increments remain stable or

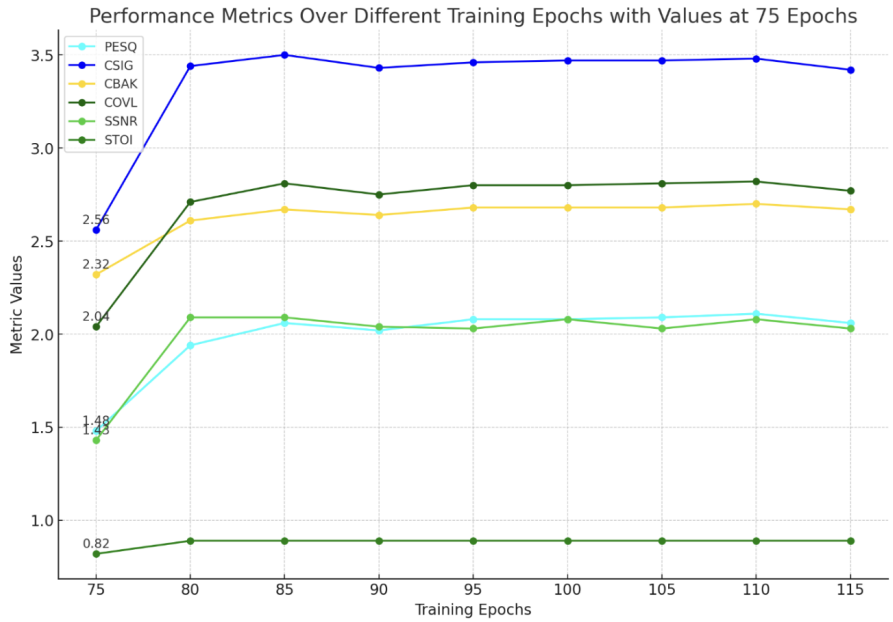


Figure 8: performance of Fine-tuned Model

Pre-trained model	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Baseline (75 epochs)	**1.48**	**2.56**	**2.32**	**2.04**	**1.43**	**0.82**
80 epochs	1.94	3.44	2.61	2.71	2.09	0.89
85 epochs	2.06	3.50	2.67	2.81	2.09	0.89
90 epochs	2.02	3.43	2.64	2.75	2.04	0.89
95 epochs	2.08	3.46	2.68	2.80	2.03	0.89
100 epochs	2.08	3.47	2.68	2.80	2.08	0.89
105 epochs	2.09	3.47	2.68	2.81	2.03	0.89
110 epochs	2.11	3.48	2.70	2.82	2.08	0.89
115 epochs	2.06	3.42	2.67	2.77	2.03	0.89

Figure 9: Detailed Evaluation Metrics of Fine-tuned Model

slightly declined. The peak scores achieved were PESQ at 2.11, CSIG at 3.50, CBAK at 2.70, COVL at 3.12, SSNR at 2.09, and STOI at 0.89. These results validate the hypothesis that fine-tuning a pre-trained model on a domain-specific dataset can significantly enhance performance, particularly in terms of perceptual quality and noise handling.

5.4 Experiment 4 Result: Comparative Analysis

In Experiment 4, the performance of the chosen model (model_21 epochs from Experiment 2) was evaluated across 20 subsets, each representing different in-car noise conditions at varying signal-to-noise ratios (SNRs) of 15 dB, 10 dB, 5 dB, and 0 dB. The model's performance was tested against five types of noise: Type1 (driving noise), Type2 (car horn), Type3 (engine idling), Type4 (air conditioner), and Type5 (street noise). These results are organized according to noise type and SNR to facilitate a direct comparison across different conditions.

Based on the observation from the table 10 and figure 11, we can see a general trend across all metrics. As the SNR increases (the higher SNR value, the stronger the signal compared to the background noise), there is a consistent increase across all metrics scores. In this experiment, this trend demonstrated that the SE model performs better when the noise level is relatively lower compared to the speech signal, since the speech is clearer and more prominent in the mixed noisy audio when the SNR value is higher.

AS for the performance by noise type demonstrated in the figure 11, we can see the multiple patterns that the SE model has under different noise conditions.

Type1 (Driving Noise): This condition showed the highest enhancement scores across almost all metrics, especially at lower SNR levels. This suggests that the model is particularly effective at mitigating continuous driving noises even when they are intense.

Type2 (Car Horn) and Type4 (Air Conditioner): These noise types showed moderate improvement patterns. Their performance curves are quite similar, possibly due to the aperiodic nature of car horns and the steady-state noise of air conditioners, which might be handled similarly by the model.

Type3 (Engine Idling) and Type5 (Street Noise): Both types showed the least improvement, particularly at 0 dB SNR, where the model struggled significantly. This indicates a lower effectiveness of the model in conditions where the noise has complex patterns or a broad frequency spectrum. And the engine idling and street noise can be considered as the most challenging noise types when conducting SE in the automotive environment.

Besides, focusing on the metric-specific outcomes, we can observe that the metrics PESQ and STOI, which respectively measure the perceptual quality of speech and its intelligibility, exhibited noticeable improvements as the SNR increased. This indicates that the model effectively enhances both the quality and comprehensibility of speech in noisy environments. Conversely, while the CSIG, CBAK, and COVL metrics also showed improvements, which were milder compared to PESQ and STOI. This suggests that while the speech becomes clearer and more intelligible, some noise elements may still be perceptible in the background.

In terms of comparative insights, the model demonstrated superior performance under driving noise conditions (Type1) as opposed to street noise (Type5) and engine idling (Type3). The potential reason behind this may be that the model is better at handling continuous, non-fluctuating noises rather than intermittent or complex noise patterns.

Comparative Analysis Based on Noise Conditions

Noise Type	SNR (dB)	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Type1 (driving_noise)	0	2.24	3.72	2.80	3.02	2.23	0.88
	5	2.39	3.91	2.88	3.19	2.42	0.89
	10	2.48	4.02	2.94	3.29	2.58	0.90
	15	2.48	4.07	2.95	3.32	2.76	0.90
Type2 (car_horn)	0	2.00	3.46	2.66	2.76	1.94	0.85
	5	2.28	3.78	2.82	3.07	2.26	0.87
	10	2.42	3.90	2.90	3.20	2.45	0.88
	15	2.53	4.05	2.96	3.33	2.66	0.89
Type3 (engine idling)	0	1.90	3.27	2.61	2.62	1.96	0.84
	5	2.16	3.61	2.76	2.93	2.28	0.86
	10	2.32	3.79	2.85	3.09	2.46	0.88
	15	2.47	3.97	2.94	3.26	2.72	0.89
Type4 (air_conditioner)	0	1.98	3.52	2.66	2.79	2.06	0.85
	5	2.21	3.77	2.79	3.03	2.32	0.87
	10	2.43	3.98	2.90	3.24	2.45	0.88
	15	2.55	4.12	2.97	3.38	2.64	0.89
Type5 (street_noise)	0	1.77	3.23	2.55	2.53	1.94	0.82
	5	2.01	3.50	2.68	2.79	2.21	0.85
	10	2.22	3.74	2.80	3.02	2.51	0.87
	15	2.40	3.95	2.90	3.21	2.70	0.88

Figure 10: Detailed Metrics of Comparative Analysis

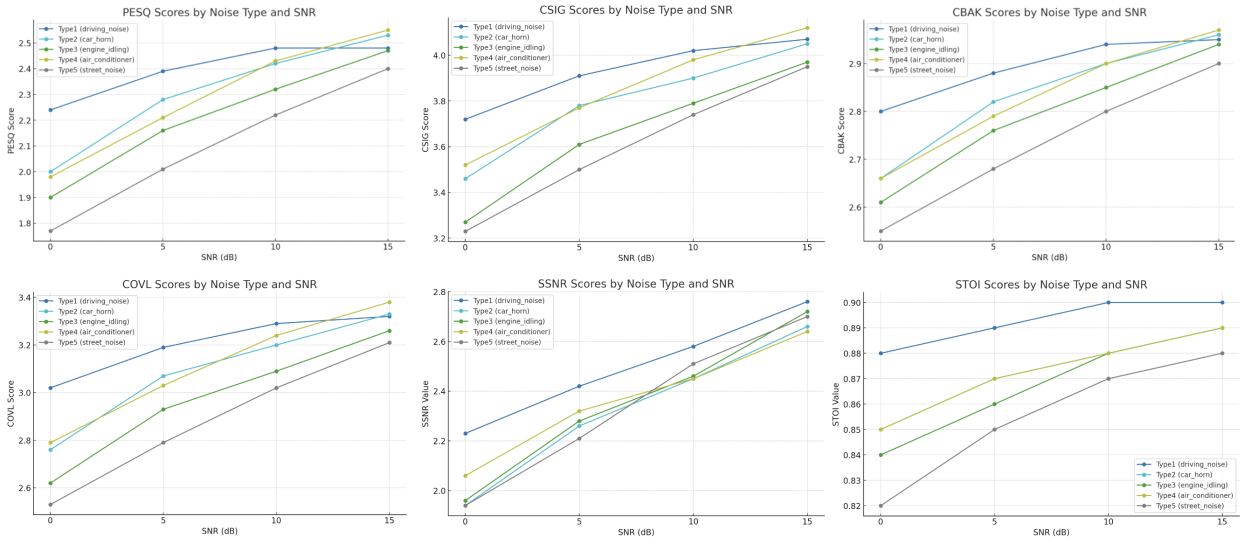


Figure 11: Evaluation Scores Comparison on 5 Noise Types

6 Discussion

After interpretation and analysis of the experiment results in the previous part, this chapter will discuss the insights and key takeaways derived from the four experiments conducted as part of this research on the CMGAN model for speech enhancement in automotive cabin noisy environments. Each experiment was designed to test specific hypotheses about the model's performance across various noise conditions and training strategies. The outcomes not only provide a detailed understanding of the model's capabilities but also highlight potential areas for refinement and further research. By dissecting the performance of the CMGAN model under different conditions, we gain valuable perspectives on optimizing speech enhancement technologies for practical applications. Here, we critically analyze the results of each experiment, discuss their significance in the broader context of speech enhancement technology, and explore the practical applications of these findings in enhancing automatic speech recognition systems in multiple noisy settings.

6.1 Discussion

For experiment 1, this initial experiment revealed that models trained on general noise types underperform when tasked with specific noise reduction challenges. This underscores a critical gap in our current approach to SE models, which tend to generalize rather than specialize. The results highlight the necessity of developing specialized SE models that are tailored to specific types of noise. This approach could potentially offer more precise enhancements in certain SE tasks where noise conditions are not universally predictable but are instead environment-dependent.

The second experiment identified an optimal checkpoint that serves as a foundation for further comparative analyses across varied noise types. Notably, even when compared to a fine-tuned model, the performance of this checkpoint from Experiment 2 consistently excelled, underscoring the efficacy of dedicating sufficient time and resources to training models specifically for distinct noise environments. This finding suggests that targeted training, while resource-intensive, yields superior results in speech enhancement tasks, particularly in specialized noisy settings.

Experiment 3 validated the hypothesis that fine-tuning the CMGAN model on a customized VCTK dataset, initially built upon a pre-trained model, enhances performance across a broad range of evaluation metrics. The success of this approach demonstrates an efficient strategy for model development under specific noise conditions, balancing time and cost considerations. The hybrid training method, where the model first learns general noise characteristics before being fine-tuned on a specialized dataset, proves particularly effective. This technique allows for the model to adapt to the unique challenges presented by specific noise environments without the need for extensive computational resources typically required for training from scratch.

Further investigation in Experiment 4 detailed how different noise types interact with the speech enhancement algorithm's performance. The results indicated that the CMGAN model performs exceptionally well with consistent noise types such as driving noise, achieving high scores across all metrics. However, the model struggled with more complex noise types like street noise, suggesting a limitation in its current capability to handle unpredictable noise patterns. The evaluation metrics used, including PESQ and STOI, showed significant improvement as the signal-to-noise ratio (SNR) increased within the same noise type class. This trend highlights the model's effectiveness in enhancing speech intelligibility and perceptual quality in controlled noise conditions.

The discussions from these experiments collectively suggest that while the CMGAN model offers

promising advancements in handling some types of noise, its performance varies significantly across different acoustic environments. The model's strengths in dealing with continuous, non-fluctuating noises provide a strong basis for further development. However, its lesser effectiveness against complex, variable noises like those found in street noise subsets indicates the need for continued research and adaptation of the model architecture or training strategies to cover a wider range of real-world noise conditions.

This discussion section not only elaborates on the findings from each experiment but also connects these results to practical implications and future research directions, providing a deeper understanding of the CMGAN model's capabilities and limitations in diverse noise environments.

6.2 Limitations

Nevertheless, our study still faces several limitations that impact the generalizability and applicability of our findings. Firstly, the custom in-car speech dataset employed in our research includes only five types of noises across four signal-to-noise ratio (SNR) levels, resulting in a total of 20 distinct noise conditions. This is just a very preliminary simplified simulation of the in-car noise environment. Previous research by (Hou et al., 2011) underscores that the acoustical environment inside a vehicle is highly complex and dynamic, influenced by factors such as driving speed, road conditions, weather, and the vehicle's soundproofing qualities. These real-world factors create a variability that our artificial noise signals fail to fully capture.

Moreover, the placement and technology of microphone arrays play a crucial role in capturing audio within such a complex environment. Far-field speech recognition, which is integral to realizing natural human-machine interaction in vehicles, requires sophisticated signal processing techniques to address the challenges posed by multiple, distant microphones being placed in the corners or around chairs. This is particularly problematic in scenarios where ambient noise levels are high and speech signals are weak or distorted by the vehicle's internal acoustics.

The challenges of Distant Automatic Speech Recognition (DASR) are well documented, given the adverse effects of room reverberation, external noise, and the dynamic nature of interactions (e.g., speaker movements). These factors significantly degrade the quality of captured speech, making speech recognition and subsequent processing arduous. Our research relies on speech enhancement algorithms that, while beneficial, are insufficient on their own for overcoming the difficulties associated with far-field speech recognition in noisy, reverberant environments. These algorithms typically enhance speech intelligibility and quality but do not adequately address the complex interference and noise typical of in-car environments.

Furthermore, the ambient scenario assumes that speakers do not modify their speech articulation to accommodate a machine listener, which is often not the case in practical applications such as voice-activated controls or dictation systems. This discrepancy between ideal conditions and real-world usage further complicates the application of our findings. In conclusion, while our study advances the field of in-car speech recognition by highlighting the potential of using microphone arrays and speech enhancement techniques, it also underscores the need for more robust, adaptive solutions that can handle the full spectrum of real-world driving conditions and user behaviors. Future research should focus on developing algorithms that can adapt more effectively to the complexities of the in-car acoustic environment, possibly through the integration of machine learning techniques that can learn from diverse and unpredictable environmental data.

6.3 Future Research

While the current results are promising, several directions remain open for future research. Firstly, due to the time limit, this study only used the evaluation metrics that were generated to test the model's performance. However, conducting a subjective evaluation study with real listening tests would provide deeper insights into the perceived quality of enhanced speech by human listeners. Such studies are important as they often reveal user-centric strengths and weaknesses not apparent through objective metrics alone. This could be considered one of the limitations of this preliminary research.

Additionally, expanding our research to include other speech enhancement tasks like dereverberation and audio superresolution would demonstrate the adaptability of the CMGAN framework to a broader range of audio processing challenges. Dereverberation, which is a common issue in many real-world environments, would test CMGAN's ability to handle echo and reverberation in audio signals, while audio superresolution would explore the model's capacity to reconstruct high-resolution audio details from low-resolution inputs.

Moreover, exploring the implementation of the CMGAN framework in real-time applications would be another good application of the SE model. Real-time speech enhancement has immense applications in telecommunications and consumer electronics, such as in smartphones and voice assistants. To the best of the knowledge, adapting GAN-based models to operate efficiently in real-time scenarios are still facing significant technical challenges due to computational constraints and the need for instant audio processing without perceptible delays.

These future directions would not only broaden the applicability of CMGAN but also contribute to the field of speech processing by providing a more holistic approach to solving diverse audio enhancement problems. Through these expansions, we can continue to push the boundaries of what is possible in speech enhancement technologies.

7 Conclusion

The primary goal of this research is to find an effective model architecture to conduct the speech enhancement on the specific noisy environment, automotive cabin. After going through the literature recently that involves all kinds of neural networks, this study focused on the development and optimization of the generative adversarial networks. In this research, we introduced the CMGAN model for speech enhancement tasks. This model operates uniquely on both magnitude and complex spectrogram components. The integration of recent conformers within CMGAN is important as it enables the capture of both long-term dependencies and localized features across time and frequency dimensions.

By focusing on the development and optimization of the CMGAN model, this study explored innovative approaches to train and fine-tune ASR technologies specifically for challenging acoustic conditions. The main method employed involved a combination of pre-training processes on a general dataset followed by targeted fine-tuning on noise-specific datasets. This hybrid training strategy was tested through a series of experiments designed to assess the model's efficacy across various noise types and signal-to-noise ratios.

The main finding of this study is that specific training on tailored datasets significantly improves the ASR system's ability to perform in noisy conditions compared to models trained on generalized noise data. This result was consistently demonstrated across multiple experimental setups, where the CMGAN model showcased enhanced speech intelligibility and quality in particularly challenging noise environments. Notably, the model excelled in conditions similar to those it was specifically fine-tuned for, underscoring the effectiveness of specialized training.

To sum up, this research demonstrated the feasibility and effectiveness of using GANs for speech enhancement tailored to specific noise types. The study's approach to training and optimization not only addresses the noise problem of ASR systems but also provides a repeatable methodology for adapting these systems to varied real-world environments. Furthermore, the insights gained from this research highlight the potential for future advancements in ASR technology, suggesting that continued refinement of training strategies and model architectures could lead to even more robust and adaptive speech recognition systems. This work contributes to the broader field of speech processing by providing a clear methodology for enhancing ASR accuracy in noisy settings, thereby improving user experience in a range of applications from automotive systems to public address systems and smart home devices.

References

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, *27*(2), 113–120.

Cao, R., Abdulatif, S., & Yang, B. (2022, September). CMGAN: Conformer-based Metric GAN for Speech Enhancement. In *Interspeech 2022.* ISCA.

Chaudhari, A., & Dhonde, S. B. (2015, January). A review on speech enhancement techniques. In *2015 International Conference on Pervasive Computing (ICPC).* Pune, India: IEEE.

Chu, P., & Messerschmitt, D. (1982). A frequency weighted itakura-saito spectral distance measure. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *30*(4), 545–560.

Donahue, C., Li, B., & Prabhavalkar, R. (2018, April). Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Fu, S.-W., Liao, C.-F., Tsao, Y., & Lin, S.-D. (2019, May). *MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement.* arXiv.

Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., & Kawai, H. (2018, March). *End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks.* arXiv.

Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., & Tsao, Y. (2021, June). *MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement.* arXiv.

Ghosh, A., Kumar, H., & Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Hansen, J. H., & Pellom, B. L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *Icslp* (Vol. 7, pp. 2819–2822).

Hao, X., Shan, C., Xu, Y., Sun, S., & Xie, L. (2019). An attention-based neural network approach for single channel speech enhancement. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 6895-6899). doi: 10.1109/ICASSP .2019.8683169

Hou, J., Liu, Y., Zhang, C., & Huang, S. (2011, November). An In-car Chinese Noise Corpus for Speech Recognition. In *2011 International Conference on Asian Language Processing.* Penang, Malaysia: IEEE.

Hu, Y., & Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, *16*(1), 229–238.

Kim, C., & Stern, R. M. (2009). Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In *tenth annual conference of the international speech communication association.*

Kim, E., & Seo, H. (2021). Se-conformer: Time-domain speech enhancement using conformer. In *Interspeech* (pp. 2736–2740).

Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. In *Icassp'82. ieee international conference on acoustics, speech, and signal processing* (Vol. 7, pp. 1278–1281).

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of ieee pacific rim conference on communications computers and signal processing* (Vol. 1, pp. 125–128).

Kundu, A., Chatterjee, S., Murthy, A. S., & Sreenivas, T. (2008). Gmm based bayesian approach

to speech enhancement in signal/transform domain. In *2008 ieee international conference on acoustics, speech and signal processing* (pp. 4893–4896).

Lee, K. Y., McLaughlin, S., & Shirai, K. (1998). Speech enhancement based on neural predictive hidden markov model. *Signal Processing*, *65*(3), 373–381.

Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013a). Speech enhancement based on deep denoising autoencoder. In *Interspeech* (Vol. 2013, pp. 436–440).

Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013b, August). Speech enhancement based on deep denoising autoencoder. In *Interspeech 2013*. ISCA.

Meng, Z., Li, J., Gong, Y., & Juang, B.-H. F. (2018, September). Adversarial Feature-Mapping for Speech Enhancement. In *Interspeech 2018*. ISCA.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

O'Shaughnessy, D. (2024, February). Speech Enhancement—A Review of Modern Methods. *IEEE Trans. Human-Mach. Syst.*.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Park, S. R., & Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

Pascual, S., Bonafonte, A., & Serrà, J. (2017, June). *SEGAN: Speech Enhancement Generative Adversarial Network*. arXiv.

Priyanka, S. S. (2017, April). A review on adaptive beamforming techniques for speech enhancement. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. Vellore: IEEE.

Recommendation, I. (2003). Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *ITU-T recommendation*, 835.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). 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)* (Vol. 2, pp. 749–752).

Salamon, J., Jacoby, C., & Bello, J. P. (2014, November). A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia*. Orlando Florida USA: ACM.

Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. (2016, September). Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. In *Interspeech 2016*. ISCA.

Valentini-Botinhao, C., et al. (2017). Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*.

Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. (2016). Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *Ssw* (pp. 146–152).

Wang, K., He, B., & Zhu, W.-P. (2021). Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7098–7102).

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B.

(2015). Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent variable analysis and signal separation: 12th international conference, lva/ica 2015, liberec, czech republic, august 25-28, 2015, proceedings 12* (pp. 91–99).

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press.