الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

جامعة غرداية
University of Ghardaia

كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science

Master Thesis Defense Permission
إذن بمنـــاقشة مذكرة مـــاستر

I, the undersigned, **Abderrahmane Adjila**

Hereby certify that I have examined the work entitled **Speech Denoising using Self-Supervised Learning Techniques** presented to the partial fulfillment of the Master degree in Computer Science by:

1. **Aya LAOUAR**

I reviewed the document, and declare it is free from any serious defaults and respects the academic integrity rules. Furthermore, my jury member proposal is the following:

Youcef Mahdjoub	Univ.Ghardaia	President
Asma Bouchekouf	Univ.Ghardaia	Examiner
Abderrahmane Adjila	Univ.Ghardaia	Supervisor
Slimane Bellaouar	Univ.Ghardaia	Co-Supervisor

Issued for all due intents and purposes.

Ghardaia, on September 9, 2024
Signature

الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

Registration n°:
...../...../...../...../.....

جامعة غرداية
University of Ghardaia
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science
مخبر الرياضيات والعلوم التطبيقية
Mathematics and Applied Sciences Laboratory

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer Science
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction

Topic

Speech Denoising using Self-Supervised Learning Techniques

Presented by:

Aya Laouar

Publicly defended on 22 09, 2024

Jury members:

Mr.Youcef Mahdjoub	MCB	Univ. Ghardaia	President
Mrs. Asma Bouchekouf	MCB	Univ. Ghardaia	Examiner
Mr. Abderrahmane Adjila	MAA	Univ. Ghardaia	Supervisor
Mr. Slimane Bellaouar	MCA	Univ. Ghardaia	Co-Supervisor

Academic Year: 2023/2024

Acknowledgment

First, I thank God Almighty. Thanks to him, I gained the faith and strength to accomplish this work. I would like to sincerely thank Dr. Abderrahmane Adjila and Dr. Bellaouar Slimane for their valuable advice, their continuous encouragement, and above all for their great willingness to complete this work. I thank them for all the trust they gave me. And also for the inspiration, help and time they devoted to me. Without them, this memoire would not have been such a success. To all members of the jury, you honor me greatly by accepting the verdict on my humble work, and you find here an expression of my sincere gratitude and respect. Finally, I would like to thank all those who have contributed closely or remotely to the completion of my work.

Dedication

Thanks to Almighty God, I have completed this modest work and I would like to dedicate it very warmly to: My dear mother and my father, who encouraged and supported me throughout my studies and for their patience, may God protect them and keep them for me. To my brothers and sisters. To the whole Laouar family. And to my very dear friends and comrades for all the moments of exchange and debate, to the people who have always helped and supported me.

Aya

ملخص

تتناول هذه الأطروحة مشكلة تقليل ضوضاء الكلام، والتي تمثل تحديًا رئيسيًا في معالجة الصوت. يمكن لإشارات الكلام الصاخبة أن تؤدي إلى تدهور كبير في أداء العديد من التطبيقات القائمة على الكلام، مثل التعرف على الكلام، والتواصل الصوتي، وتحسين الصوت. الفرضية الرئيسية التي يقوم عليها هذا العمل هي أنه يمكن الاستفادة بفعالية من تقنيات التعلم الذاتي الإشراف لتقليل الضوضاء في الكلام من دون الحاجة إلى مجموعات بيانات كبيرة . الأهداف الأساسية هي استكشاف إمكانيات الأساليب ذاتية الإشراف لتقليل ضوضاء الكلام، وتنفيذ وتقييم الخوارزميات التي يمكنها إزالة الضوضاء من الكلام مع الحفاظ على جودة الكلام الأساسية. يتضمن نهج البحث مراجعة الأدبيات حول الأساليب التقليدية والحديثة، وتنفيذ نموذج التعلم الذاتي الإشراف، وإعداد البيانات الصوتية ومعالجتها مسبقًا، بالإضافة إلى تدريب النموذج وتقييمه.

تشير نتائج هذا العمل إلى أن التعلم الذاتي الإشراف يمكن أن يكون اتجاهًا واعدًا لمعالجة مشكلة تقليل الضوضاء في الكلام، وخاصة في البيئات حيث يكون الحصول على البيانات المُعلقة نادرًا أو مكلفًا. يمكن دمج هذا النموذج في تطبيقات مختلفة تعتمد على الكلام لتحسين أدائها ومتانتها في البيئات الصاخبة. قد تتضمن اتجاهات البحث المستقبلية استكشاف تقنيات أكثر تقدمًا للإشراف الذاتي والتحقيق في إمكانية نقل التمثيلات المكتسبة إلى مهام معالجة الصوت الأخرى.

كلمات مفتاحية: تقليل الضوضاء من الكلام، التعلم ذاتي الإشراف، معالجة الصوت، التعلم الآلي، التعلم العميق.

Abstract

This thesis addresses the problem of speech noise reduction (NR), which represents a major challenge in audio processing. Noisy speech signals can significantly degrade the performance of many speech-based applications, such as speech recognition , voice communication and voice enhancement . The main hypothesis underlying this work is that self-supervised learning (SSL) techniques can be effectively leveraged to reduce noise in speech without the need for large datasets. The primary goals are to explore the possibilities of SSL methods for NR, to implement and evaluate algorithms that can remove noise from speech while maintaining basic speech quality. The research approach includes reviewing literature on traditional and modern methods, implementing an SSL model, preparing and pre-processing audio data, as well as training and evaluating the model. The results of this work suggest that SSL could be a promising direction to address NR in speech, especially in environments where obtaining annotated data is rare or expensive. This model can be integrated into various speech-based applications to enhance their performance and robustness in noisy environments. Future research directions may include exploring more advanced SSL techniques and investigating the transferability of learned representations to other sound processing tasks.

Keywords: Speech denoising, Self-Supervised Learning, Audio processing, Machine learning, Deep learning.

Résumé

Cette thèse aborde le problème de la réduction du bruit de la parole, qui représente un défi majeur dans le traitement audio. Les signaux de parole bruyants peuvent dégrader considérablement les performances de nombreuses applications basées sur la parole, telles que la reconnaissance vocale, la communication vocale et l'amélioration de la voix. L'hypothèse principale sous-jacente à ce travail est que les techniques d'apprentissage auto-supervisé peuvent être efficacement exploitées pour réduire le bruit de la parole sans avoir besoin de grands ensembles de données . Les principaux objectifs sont d'explorer les possibilités des méthodes auto-supervisées pour la réduction du bruit de la parole, de mettre en œuvre et d'évaluer des algorithmes capables de supprimer le bruit de la parole tout en maintenant la qualité de base de la parole. L'approche de recherche comprend l'examen de la littérature sur les méthodes traditionnelles et modernes, la mise en œuvre d'un modèle d'apprentissage auto-supervisé, la préparation et le prétraitement des données audio, ainsi que la formation et l'évaluation du modèle. Les résultats de ce travail suggèrent que l'apprentissage auto-supervisé pourrait être une direction prometteuse pour résoudre le problème de la réduction du bruit de la parole, en particulier dans les environnements où l'obtention de données annotées est rare ou coûteuse. Ce modèle peut être intégré dans diverses applications basées sur la parole pour améliorer leurs performances et leur robustesse dans les environnements bruyants. Les futures orientations de recherche pourraient inclure l'exploration de techniques auto-supervisées plus avancées et l'étude de la transférabilité des représentations apprises à d'autres tâches de traitement du son.

Mots clés: Débruitage de la parole, apprentissage auto-supervisé, traitement audio, apprentissage automatique, apprentissage profond.

Contents

List of Figures

Introduction

Speech recognition and processing are very active areas of research in computer science and signal processing, having experienced significant advances in recent years, notably due to developments in machine learning and deep learning. However, one of the major challenges remains speech denoising, or the removal of noise present in speech signals. Noise can come from various sources such as the acoustic environment, electronic interference, or sensor imperfections, all of which significantly degrade speech quality and intelligibility. This degradation can impact many applications, such as speech recognition, telephony, audio streaming, and others. Therefore, the development of effective speech denoising methods is a crucial task.

Traditional denoising techniques, such as statistical approaches and time-domain and spectral-domain methods, often rely on simplifying assumptions about the statistical properties of the signal and noise. These techniques, however, have limitations, especially when the noise is non-stationary, colored, or strongly correlated with the speech signal. In this context, machine learning approaches, particularly deep learning methods, have shown significant potential in learning more flexible and adapted models for noisy speech signals. Some examples of machine learning techniques applied to speech denoising include supervised learning, which uses paired clean and noisy speech to train enhancement models, as well as unsupervised and self-supervised approaches that learn meaningful representations directly from noisy data without needing clean speech references.

Self-supervised learning (SSL) is an approach that has recently gained traction in speech denoising due to its ability to leverage large amounts of unlabelled data. The core principle of SSL is to create pretext tasks from the raw data itself, where the model learns to predict a certain part or aspect of the data based on other parts, thus creating “labels” from the data automatically. For example, in speech denoising, an SSL model might be trained to predict a clean version of a speech segment given a noisy version. Through this process, the model learns representations of speech that capture the essential features required for denoising without needing manually annotated clean-noise pairs. This approach is particularly beneficial because it can adapt to real-world noise conditions where labeled data is costly or difficult to obtain.

This thesis aims to explore the use of self-supervised learning techniques for speech denoising. SSL enables the use of large amounts of unlabelled audio data to learn speech signal representations that capture the most relevant features for the denoising task without requiring costly manual annotations. The objective of this work is to study and evaluate self-supervised learning models for speech denoising, improving the quality and usability of speech-based applications in real-world environments. The thesis will be organized into three chapters. The first will cover general concepts about speech, noise, and machine learning methods. The second chapter will review in detail various speech noise removal methods and the latest findings in self-supervised speech noise reduction techniques. The third part will focus on a study and implementation of the proposed model for self-supervised speech noise reduction. Finally, the conclusion will summarize the main achievements and outline directions for future research.

Chapter 1

Generalities about speech and noise.

1.1 Introduction

Speech is one of the primary means of communication among human beings. Its simplicity makes it the most popular communication method in human society (it is easier to speak to someone than to write or draw a diagram for them) . The chapter provides an introduction to the generalities of speech and noise. The chapter begins by discussing the importance of digital speech processing in various fields and highlights the common problem of corruption of speech signals due to various types of noise. The chapter covers several topics, including properties of audio signals, such as frequency, amplitude, and waveform, as well as specific properties of speech signals, such as acoustics and mechanisms of speech production. It also explores different techniques for visualizing audio signals and explains the process of converting analog audio signals into digital form. The chapter also delves into the types of noise that can corrupt speech signals, the concept of signal-to-noise ratio (SNR), and the challenges involved in reducing noise from speech signals. In addition, it introduces different learning methods used in speech noise reduction, including supervised learning, unsupervised learning, and self-supervised learning. Overall, this chapter serves as a foundation for understanding the concepts and challenges of speech noise reduction using self-supervised learning techniques.

1.2 Audio in General

Audio is a form of communication that relies on acoustic waves to convey information. It encompasses various elements such as speech, music, and sound effects. The use of audio is widespread in daily life, from watching movies and listening to music to attending lectures and making phone calls. One key aspect of audio is acoustic waves, vibrations that travel through a medium like air or water. These waves can be described by their frequency, determining pitch, and amplitude, determining volume. The human ear's ability to perceive audio is crucial, consisting of three main parts: the outer ear, middle ear, and inner ear. The outer ear collects audio waves and directs them to the middle ear, where they are amplified and transmitted to the inner ear. In the inner ear, audio waves are converted into electrical signals processed by the brain. This process is essential for understanding and interpreting auditory information. In conclusion, audio is a fundamental aspect of communication and entertainment, involving the transmission and perception of acoustic waves and relying on various techniques to create different sounds (Ballas (2007)).

1.2.1 Characteristics of Audio

When we talk about sound , we are referring to the physical phenomenon that relates to the transmission of mechanical disturbances through a material medium, such as air or water. Sound is produced by generating mechanical disturbances that are transformed into sound waves that propagate in the surrounding medium. Sound has several characteristics, including:

Sound Waves Sound waves are longitudinal waves of pressure variations transmitted through a medium, such as air, water, or solids. These variations create compressions and rarefactions, causing particles in the medium to move back and forth. Sound waves are the physical manifestation of vibrations of Music (2024).

$$y(x,t) = A\sin(2\pi ft - \Phi) \tag{1.1}$$

Where: **y** is displacemment , $\boldsymbol{A}$ is amplitude , $\boldsymbol{f}$ is the sound frequency ,$\boldsymbol{t}$ is time, $\boldsymbol{x}$ is position and **Φ** is the phase angle.

Frequency Frequency is the number of cycles of a sound wave that occur in one second, measured in Hertz (Hz). It determines the pitch of the sound. Higher frequencies result in higher-pitched sounds, while lower frequencies produce lower-pitched sounds. The audible range for humans is typically 20 Hz to 20,000 Hz OpenStax (2024).

$$f = \frac{1}{T} \tag{1.2}$$

Where $\boldsymbol{f}$ is frequency, and $\boldsymbol{T}$ is the time period.

Amplitude Amplitude refers to the magnitude or intensity of a sound wave, determining its loudness. It is measured in decibels (dB), and a higher amplitude corresponds to a louder sound. Amplitude is a crucial factor in the perception of a sound's volume OpenStax (2024).

Duration Duration is the length of time a sound persists. It is a fundamental aspect in music, speech, and various audio applications. Sounds can be short and transient, like a drum hit, or sustained over a more extended period, such as a musical note or a spoken word.the equation of Duration is inverse of frequency of Music (2024).

$$T = \frac{1}{f} \tag{1.3}$$

The time period **T** of a sound is inversely related to its frequency

Pitch Pitch refers to the perceived frequency of a sound wave. It is commonly described as high or low and is determined by the rate at which the sound wave vibrates. Higher frequencies result in higher pitches, while lower frequencies produce lower pitches((*Audio Definition*, 2022)).

$$H = k \cdot \log_2 \left(\frac{f}{f_0} \right) \tag{1.4}$$

Where $\boldsymbol{f}$ is the current frequency, $\boldsymbol{f0}$ is the reference frequency, and $\boldsymbol{k}$ is a constant. The value of the constant k in the pitch equation 1.4depends on the specific context and the chosen reference frequency (f_0). For human speech perception, k is often determined empirically to match the subjective sense of pitch for the average human ear. A common reference frequency is 1000 Hz.

In some contexts, k might be set to a value around 3.5 to 4 for speech-related studies. However, it's important to note that the exact value can vary, and researchers might adjust it based on their specific experiments or applications. Experimentation and validation with human subjects are typical approaches to fine-tune such constants in the context of studying human sound perception.

Timbre Timbre refers to the unique quality or tone color of a sound. It distinguishes one sound from another, even if they have the same pitch and loudness. Timbre is influenced by the complex mixture of frequencies and harmonics present in a sound wave Academy (2024).

Dynamic Range Dynamic range is the difference between the softest and loudest parts of an audio signal. It is often expressed in decibels and relates to the signal-to-noise ratio.

$$DR(dB) = (20 \cdot \log_{10} \left(\frac{V_{\max}}{V_{\min}} \right)) \tag{1.5}$$

where $V_{\max}$ and $V_{\min}$ are the maximum and minimum signal voltages embibe 2023 (2023).

1.2.2 Speech

Speech is a series of sound waves generated through the oscillation of air molecules. This process is initiated when an individual expels air from their lungs, modulating the resulting sounds using the structures of the mouth and nose. These sound waves propagate through a medium and exhibit distinctive properties, including frequency which determines pitch, and amplitude which signifies the intensity of vibrations. The human ear serves to detect and interpret these waves, and a comprehensive grasp of these physical characteristics is integral to the fundamental aspects of auditory perception and communication (Lee et al., 2021).

In more straightforward terms, spoken language involves the creation of sound waves by expelling air through the vocal cords and manipulating these waves using the oral and nasal structures. These waves possess specific attributes such as pitch and volume, which are then apprehended and deciphered by our ears, facilitating effective hearing and communication McLoughlin (2016).

Characteristics

The intricate characteristics of speech form a multifaceted tapestry, encompassing both temporal and frequency domains, providing a robust foundation for analysis in diverse applications. In the temporal domain, the rhythmic patterns of speech, including phoneme duration, silent intervals, segmental timing, and prosody, intricately contribute to the emotional expression of communication. These temporal nuances are crucial components in understanding the dynamics of spoken language. On the other hand, the frequency domain of speech 1.1 reveals a rich composition through spectral content, formants, harmonics, and prosodic cues. Formants, specifically resonant frequencies in the vocal tract, stand out for their pivotal role in differentiating phonemes. This highlights their significance in the intricate structure of speech sounds. Within the physical components of speech, elements such as pitch contours, airflow and pitch rates play indispensable roles. Spectral plots and pitch lag analysis serve as tools to demonstrate and analyze these physical attributes, offering insights into the acoustic features of speech production. The amplitude distribution of speech varies dynamically based on situational factors, ranging from the subtlety of a whisper to the intensity of shouting. This variation underscores the adaptability of speech to different environmental and emotional contexts. Turning attention to the lexical components, including phoneme sequences, tone, timbre, and amplitude, contributes significantly to our understanding and interpretation of spoken language. The frequency distribution of speech closely aligns with the sensitivity of the human ear, emphasizing a notable distinction between frequencies with the greatest energy and those essential for intelligibility(McLoughlin, 2016). In the temporal dimension, speech exhibits constraints on articulation speed, with phoneme duration and syllabic rate remaining relatively constant. This stability in temporal characteristics ensures a consistent framework for speech communication. In conclusion, the exploration of speech characteristics in both the temporal and frequency domains provides a comprehensive understanding that serves as a solid foundation for further analysis and processing in diverse applications.

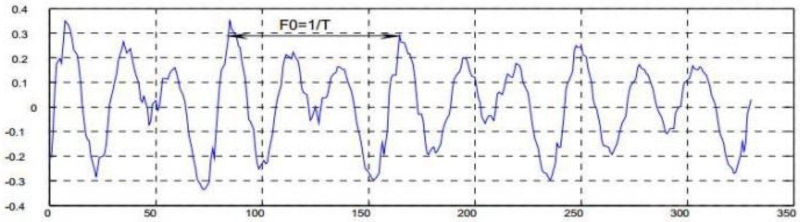


Figure 1.1: A voiced sound with its fundamental frequency

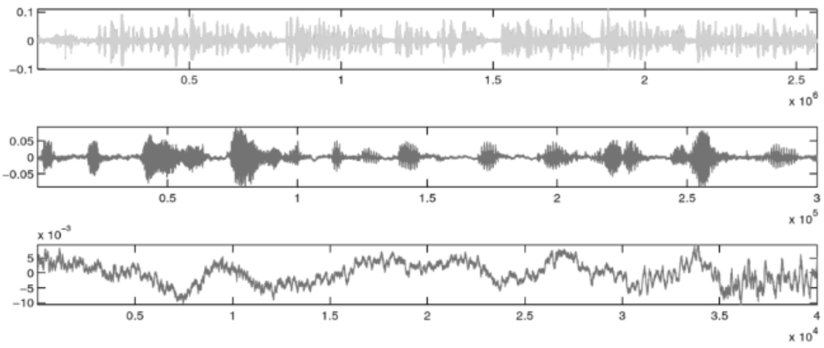


Figure 1.2: Amplitude against time plots of the same speech recording at three different time scales.

Visualization of Sounds

Exploring the concept of visualizing sound underscores its pivotal role in converting signals into visual representations, facilitating human comprehension. While this process enhances our understanding of signal complexities, it acknowledges that visualizations can potentially obscure certain aspects. Despite the sensory disparities between eyes and ears, both possessing unique strengths and weaknesses, their collaborative use in signal analysis forms the foundation for subsequent discussions.

Oscilloscope The oscilloscope, one of the oldest representations, shows the temporal evolution of the signal's amplitude. It is a simple function of time that does not reveal the internal structure of the sound (its frequency composition) and proves less interesting for complex auditory objects, especially in the study of speech. An auditory object is generally defined using three main parameters (acoustic trivariance): its intensity, its frequency composition, and its duration. Visualizing a sound involves finding a representation that is related to these three parameters, thus requiring a three-dimensional space (amplitude, frequency, time). A simple illustration of the problems inherent with a Oscilloscope(waveform) view is given in Figure1.2 , where three different resolution views of the same signal (conversational speech) reveal very little visual similarity for what is really a fairly uniform audio signal Tektronix (2024).

Frequency Spectrum Frequency Spectrum is a fundamental method for analyzing signals, particularly in understanding the distribution of energy across different frequencies within a given time frame or analysis window. This tool is crucial for capturing a snapshot of the primary frequency components present in the signal,

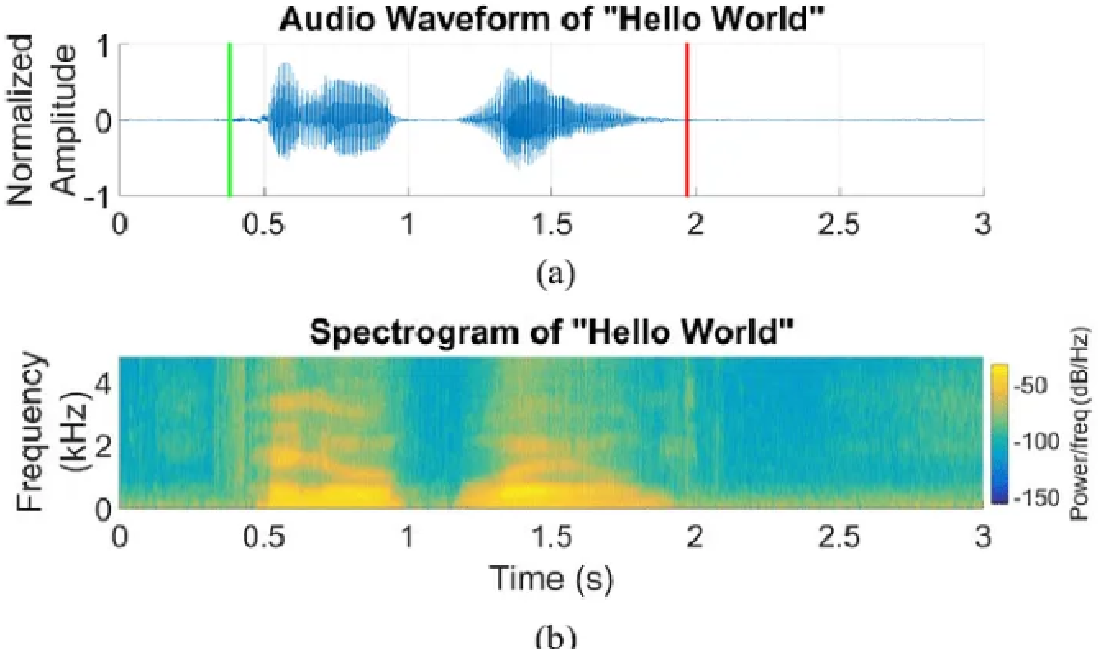


Figure 1.3: Audio waveform superimposed

offering valuable insights into its characteristics. A key aspect emphasized is the careful selection of the analysis window, as it plays a critical role in the accuracy and meaningfulness of the frequency spectrum analysis (Anonymous, 2016). The cautionary note underscores the importance of avoiding the oversight of significant features within the signal, ensuring that the chosen window aligns with the characteristics of the signal under examination.Some audio researchers prefer to plot their spectrograms in colour, but this is really just a matter of personal preference Tektronix (2024). view is given in Figure1.3

Short-Time Fourier Transform (STFT) The Short-Time Fourier Transform (STFT) stands as an advanced signal processing technique, involving the application of a Fourier transform to localized segments or windows of a signal. This process yields a time-varying representation of the signal's frequency content, proving particularly valuable when dealing with signals exhibiting variations over time. By systematically applying a narrow Fourier transform to successive windows along the signal, STFT generates a time sequence of spectra. Typically represented as a spectrogram, this visualization effectively illustrates how the frequency components of a signal evolve over distinct time intervals (Anonymous, 2016). The STFT's significance extends to audio and speech processing, where it plays a crucial role in visualizing speech structure over time. Its capability for detailed time-frequency analysis surpasses simpler methods such as the time-domain waveform plot or frequency spectrum, making it an indispensable tool in signal analysis and processing. The Short-Time Fourier Transform (STFT) is mathematically represented by the following equation:(Rocchesso, 2003)

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau) \cdot w(t-\tau) \cdot e^{-j\omega\tau}\, d\tau \tag{1.6}$$

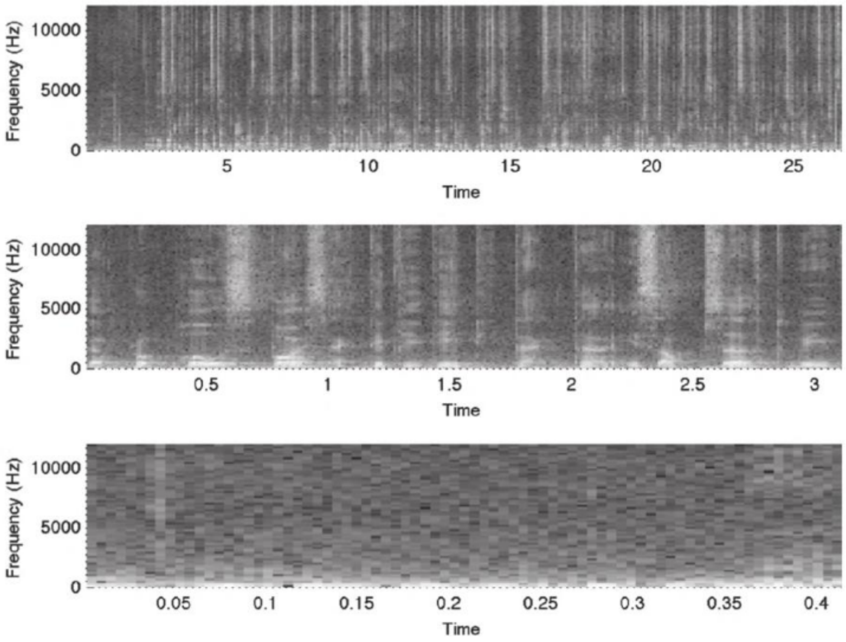


Figure 1.4: Spectrograms of the same single audio recording views as given in Figure1.2 . Louder sound components are shown with a darker grey shade and lower amplitudes with a lighter shade.

where: - $X(t, \omega)$ is the STFT of the signal. - $x(\tau)$ is the input signal. - $w(t - \tau)$ is the window function applied to the signal. - ω represents the angular frequency. - j is the imaginary unit.

An example of MATLAB spectrograms is shown in Figure1.4 , plotted in greyscale, and giving the same three periods of speech as shown in Figure 1.2

Other Visualization Methods emphasizes the significance of advanced visualization techniques, highlighting the linear prediction coefficient spectral plot, correlogram, and cepstrum. These methods are presented as tools for in-depth signal analysis, surpassing the capabilities of simpler visualization techniques. The linear prediction coefficient spectral plot provides insights into spectral characteristics, while the correlogram and cepstrum offer unique perspectives on signal autocorrelation and component separation Ignoto (2017).

Correlogram Is a visual representation of the autocorrelation of a signal. Autocorrelation involves comparing a signal with its delayed versions to identify similarities. In the context of a correlogram, this comparison is graphically presented, highlighting patterns and periodicities within the signal. MATLAB is often employed for autocorrelation analysis, aiding in the detection of hidden periodicities that may not be apparent through direct observation. Correlograms are particularly useful in speech analysis for uncovering repetitive structures within audio signals Processing (2022).

Cepstrum Derived from reversing the word ”spectrum,” is a plot of amplitude against ”quefrency,” the inverse of frequency. This technique involves a double Fourier transform process applied to the logarithm of the original signal’s Fourier transform. The cepstrum is effective in separating components within complex signals, making it valuable for speech analysis. MATLAB's signal processing toolbox provides functions like cceps() for convenient cepstral analysis. Cepstrum are commonly used to determine the fundamental frequency of a signal and have diverse applications in speech processing tasks *Mel-Cepstrum* (2021).

In conclusion, both the correlogram and cepstrum serve distinct purposes in signal analysis. Correlograms provide a visual representation of autocorrelation, aiding in the identification of periodicities and repetitive patterns. On the other hand, cepstrum excel in separating components within complex signals, offering detailed insights into signal structure. While correlograms are relatively easier to compute and scale, cepstrum are valuable for automatic peak detection and revealing features that may be missed in correlograms. The choice between these methods depends on the specific analysis goals and the nature of the signal under consideration. Both tools contribute significantly to speech analysis and other signal processing applications McLoughlin (2016).

1.2.3 Digital Representation

Digital representation of audio has undergone a transformative journey, evolving from early mechanical gramophones to the sophisticated landscape of modern electronic technology. In the historical context, audio signals were initially represented using mechanical devices like gramophones, relying on physical components to capture and reproduce sound as pressure waves. This analog representation marked the inception of audio playback but was limited by the constraints of physical mechanisms. The transition to digital representation heralded a new era, where audio signals are now encoded as discrete digital data, offering advantages in storage, processing, and transmission. This paradigm shift paved the way for enhanced fidelity, efficient data handling, and the ability to manipulate audio content in ways not achievable with analog methods. One fundamental aspect of digital representation is the encoding and decoding process, often referred to as audio coding. This process involves the conversion of analog signals into digital format for storage and subsequent reconstruction during playback. The delicate balance between factors like fidelity, data rate, complexity, and delay became crucial considerations in designing effective audio coders to cater to diverse application needs. The concept of Pulse Code Modulation (PCM) exemplifies a foundational approach to digital representation, where analog signals are sampled, quantized, and reconstructed. A prominent manifestation of PCM is the Compact Disc (CD), which became a ubiquitous medium with a sampling frequency of 44.1 kHz and 16 bits per sample, setting a standard for digital audio quality(Bosi & Goldberg, 2022). is shown in Figure 1.5

As technology advanced, more intricate coding schemes emerged, such as transform coders, which dynamically allocate bits across the frequency spectrum to achieve perceptual transparency. This evolution in digital representation has not only revolutionized audio playback but has also raised intriguing questions about the necessity of complexity, especially when considering the repetitive information inherent in certain digital audio formats. In essence, the journey of digital representation of audio signifies a profound shift from analog mechanisms to the versatile and powerful realm of digital technology, shaping the way we perceive, store, and

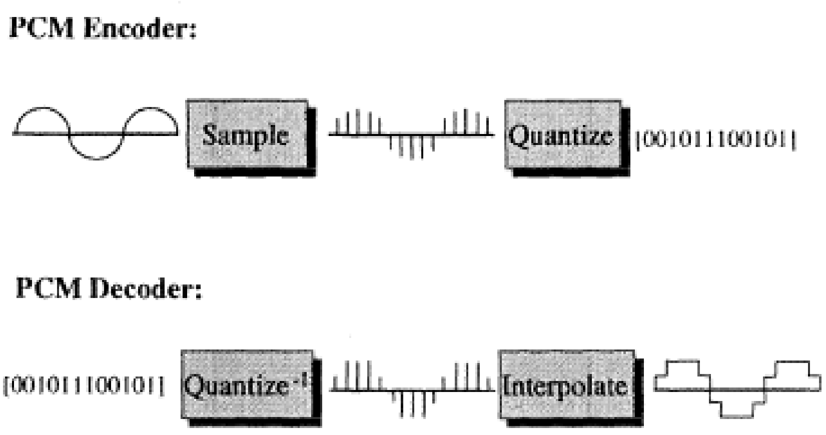


Figure 1.5: The PCM Encoder and Decoder

interact with audio content.

Basics of Sound Data Representation

Sound data representation comprises a vital multi-step process essential for digital audio processing. Initially, the continuous analog signal undergoes sampling, where discrete points are captured at regular intervals. The sampling rate, determining the frequency of these samples, plays a crucial role in reconstructing the original signal accurately. Following this, quantization comes into play, assigning a finite set of values to represent varying signal amplitudes. The quantized data is then encoded into a digital format, often binary, facilitating computer processing. Another critical aspect is bit depth, determining the precision of each sample's representation; higher bit depth enables a more faithful reproduction of the original audio. In essence, the synergy of sampling, quantization, encoding, and the specifications of sampling rate and bit depth collectively forms the foundation of digital sound representation. These processes not only facilitate the study of sound but also simplify the exploration and enhancement of audio within computational systems while minimizing noise.

Converting Analogue Audio Into Digital Sound Representation

The process of capturing and converting sound in digital audio systems involves the transformation of sound waves into electrical signals through a microphone, which detects membrane deflection caused by molecular vibrations in the air. Subsequently, an analog-to-digital converter (ADC) translates these electrical signals into coded digital data, commonly using pulse-coded modulation. Once the coded data undergoes processing, it is fed through a digital-to-analog converter (DAC) to produce sound through a loudspeaker. The voltage applied to the loudspeaker corresponds to the computer's sample values, leading to the deflection of the loudspeaker's cone and the initiation of a sound pressure wave. Key steps in digital audio processing encompass sound capture, amplification, ADC conversion, signal processing, DAC conversion, and loudspeaker output. The process of capturing and converting sound in digital audio systems involves two essential components: the Analog-to-Digital Converter (ADC) and the Digital-to-Analog Converter (DAC)(Bosi & Goldberg, 2022).

ADC (Analog-to-Digital Converter): ADC is responsible for translating analog signals, such as those captured by a microphone, into digital data suitable for processing by a computer. This process involves three key steps:

1. Sampling The process of sampling involves taking discrete points or samples of a continuous analog signal at regular intervals, known as the sampling rate, measured in samples per second (Hz). A higher sampling rate enhances the accuracy of representing the original analog signal(Rocchesso, 2003). This procedure aligns with the sampling theorem, also known as the Nyquist-Shannon theorem 1, a fundamental principle in signal processing.

Theorem 1 (Nyquist Theorem) *For lossless digitization, the sampling rate (f_s) should be at least twice the maximum frequency response ($2f_m$), where f_s is the sampling frequency and f_m is the maximum frequency in the signal.*

2. Quantization The continuous range of each sampled value is converted into a discrete set of digital values. This step involves assigning a digital code (usually in binary form) to each analog sample. The precision and dynamic range of the converted signal are influenced by the bit-depth of the ADC, which determines the number of bits in each digital code(Rocchesso, 2003).

3. Encoding After quantization, the digital values obtained from the ADC are encoded into a specific digital format. Pulse Code Modulation (PCM) is a common method used for this encoding. PCM assigns a unique binary code to each quantized amplitude value. These binary codes represent the digital equivalent of the analog signal at specific points in time. The encoded digital data can then be further processed, transmitted, or stored in various digital audio formats(Rabiner & Schafer, 2007).

DAC (Digital-to-Analog Converter) The Digital-to-Analog Converter (DAC) plays a pivotal role in the audio reproduction process, seamlessly transforming digital signals into analog signals for playback through a speaker. The intricate DAC process unfolds as follows: Upon receiving digital data, which embodies the quantized and sampled values derived from the original analog signal, the DAC initiates its transformative journey. Through the decoding phase, the DAC meticulously translates the binary codes embedded in the digital data, seamlessly restoring them to discrete digital values. The heart of the process lies in analog reconstruction, where the DAC skillfully transforms the discrete digital values into a continuous analog signal. This critical step ensures the faithful reproduction of the original waveform with precision. Subsequently, the meticulously reconstructed analog signal is channeled to a speaker or audio output device. The speaker, acting as the final emissary in this chain of transformations, faithfully translates the analog signal into sound waves that closely mirror the nuances of the initially captured sound. In

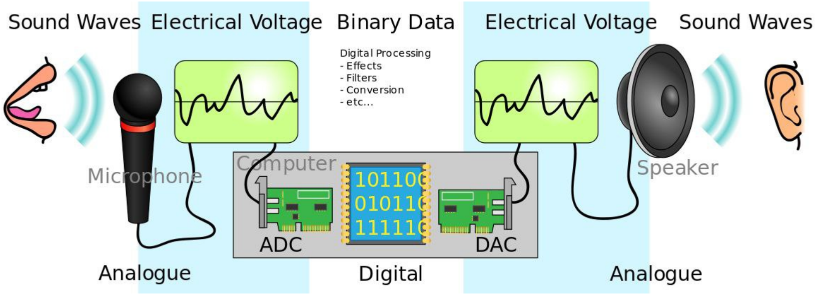


Figure 1.6: Life cycle from Sound to Digital to Sound

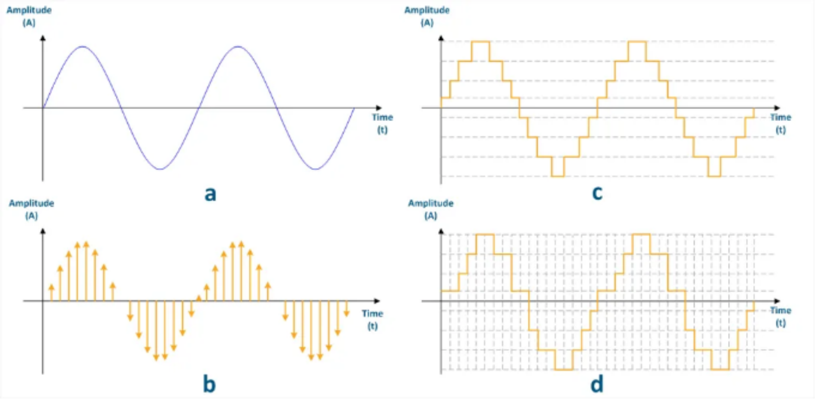


Figure 1.7: Diagram showing the flow of audio from an analog waveform to a digital binary representation,An analog wave is represented as a continuous, wavy line, while a digital signal is represented as a series of interrupted rectangular shapes. Digital binary representation is represented by a series of 0s and 1s, where the analog wave is converted into digital format.

essence, the DAC's proficiency in decoding, analog reconstruction, and delivering the final output to the speaker ensures a coherent and accurate transition from the digital realm to a perceptually rich analog soundscape.

In summary, the complete process involves the ADC capturing analog signals through sampling and quantization, followed by encoding the digital values. The DAC then decodes and reconstructs the original analog signal for output through a speaker1.6. This seamless interplay between ADC and DAC ensures the accurate and faithful conversion of analog sound into digital data and its subsequent reconstruction for human perceptionRabiner & Schafer (2007).

Importance of Digital Representation

In Enhancing Sound Quality Converting sound into digital format has revolutionized the audio industry, primarily due to the enhanced sound quality it provides. Digital representation mitigates hiss, distortion, and noise typically associated with analogue audio formats.The figure1.7 shows the representation of sound.It ensures the audio quality remains unchanged despite repeated playback or copying.It facilitates audio storage and transfer without loss of quality. Moreover, it paves the way for advanced audio processing techniques, such as equalization, noise reduction, and sound synthesis**?**.

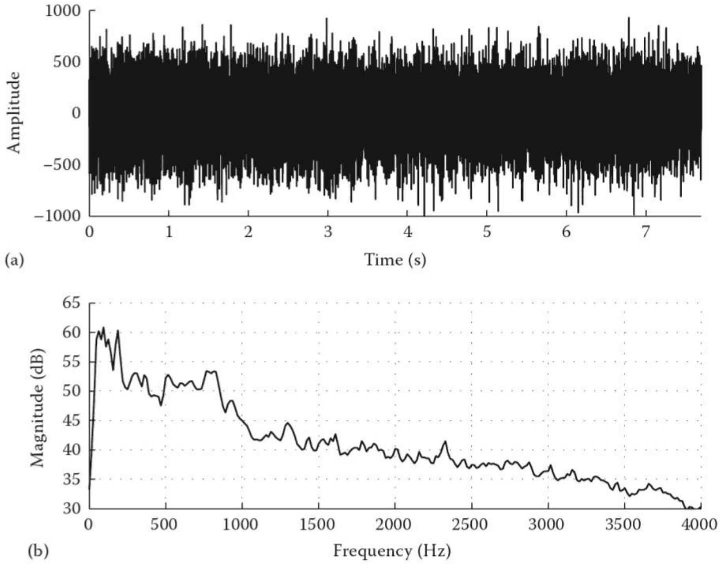


Figure 1.8: a) Example noise from a car and (b) its long-term average spectrum.

1.2.4 Noise

Noise, in the context of sound, refers to unwanted or disruptive sounds that can interfere with desired signals, often leading to a decrease in overall sound quality. Sources of noise are diverse and ubiquitous, ranging from environmental factors like traffic and wind to human activities such as machinery or conversations. The levels of noise vary significantly across different environments. For instance, in tranquil settings like classrooms or homes, noise levels tend to be lower, allowing for clearer communication. On the contrary, environments like restaurants, trains, or airplanes are characterized by higher noise levels, posing challenges for effective communication(Loizou, 2012).

Understanding and managing noise in various settings is crucial for developing strategies, including speech denoising algorithms, to mitigate its impact on communication quality. This involves recognizing different types of noise and studying their temporal and spectral characteristics. For example, wind noise predominantly concentrates on low frequencies, while restaurant noise extends over a broader frequency range. The spectral profile of noise is of utmost importance, and Figures 1.8 , 1.9 , 1.10 illustrate the time waveforms and long-term average spectra of car, train, and restaurant noise.

Noise is commonly measured in decibels (dB), which represents the ratio or attenuation of the signal being measured compared to the level of noise present. Decibels provide a standardized scale for quantifying sound intensity and are essential for assessing the impact of noise in various environments.

Types of Noises

Noise encompasses a variety of types, each uniquely characterized, playing a pivotal role in fields like signal processing and audio engineering. White Noise (contributors, 2021a), represented by random signals with equal intensity at diverse frequencies, resembles the static on a TV or radio. This type of noise can be

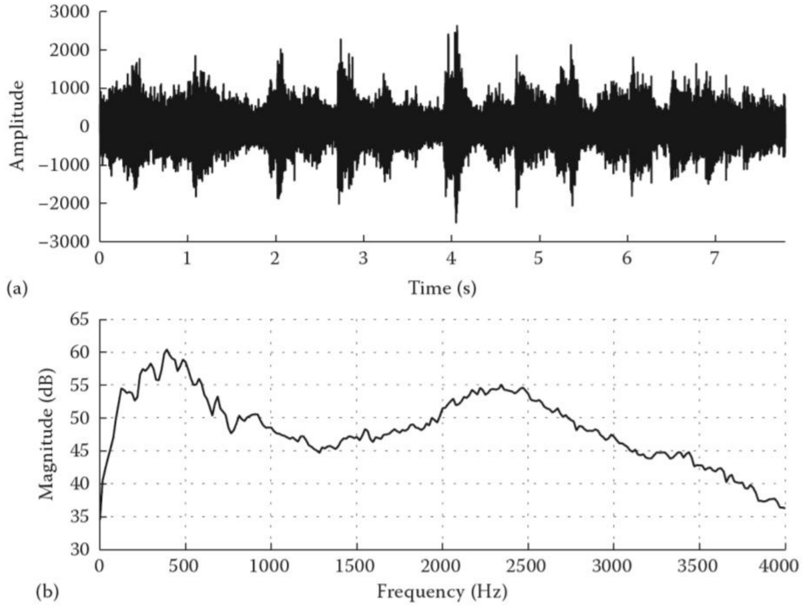


Figure 1.9: (a) Example noise from a train and (b) its long-term average spectrum.

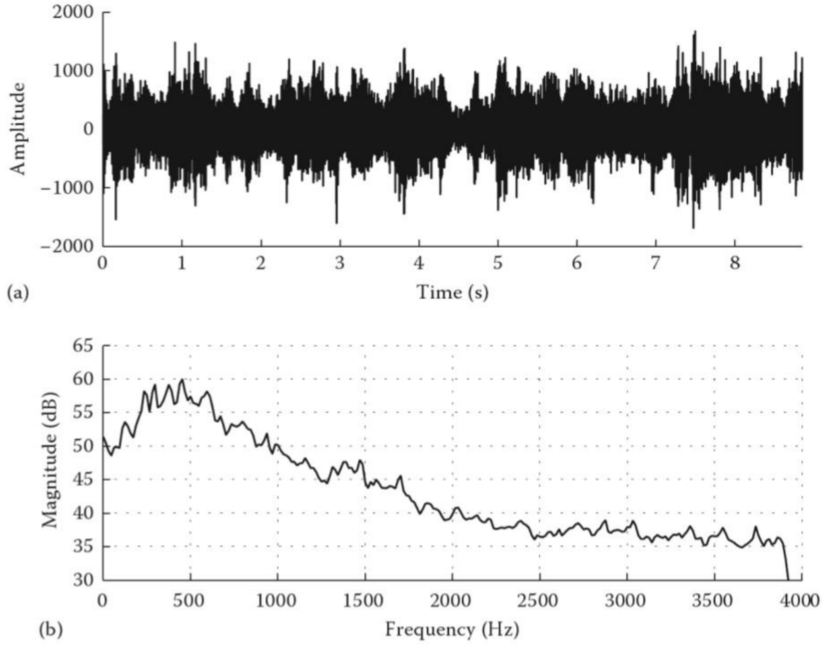


Figure 1.10: (a) Example noise from a restaurant and (b) its 1ong-term average spectrum.

mathematically expressed as

$$W(t) = A \cdot N(t) \tag{1.7}$$

where A is the amplitude and $N(t)$ is a random signal with zero mean and unit variance. Pink Noise ("Spectral Content (colour) of Noise Exposure Affects Work Efficiency", 2024) is akin to white noise but exhibits reduced intensity at higher frequencies, analogous to the soothing sounds of ocean waves. Its power spectral density at frequency f is given by

$$P(f) = \frac{A}{f^{\beta}} \tag{1.8}$$

where A is a constant, and β determines the rate of decrease with frequency. Brownian Noise (Brown Noise)(contributors, 2021b), following a 1/frequency^2 power density, showcases a decline in intensity with increasing frequency, akin to the calming ambiance of waterfalls. It can be represented as

$$B(t) = \int_0^t W(\tau)\, d\tau \tag{1.9}$$

where $B(t)$ is the Brownian noise and $W(t)$ is the white noise. Gaussian Noise (Lakin, 2013), adhering to a Gaussian distribution, embodies randomness, typified by thermal noise in electronic circuits. Its probability density function is given by

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.10}$$

where μ is the mean, and σ is the standard deviation. Impulse Noise introduces sudden, brief disturbances, like clicks in audio signals or pixel errors in images. It can be expressed as

$$I(t) = \sum_k a_k \delta(t - t_k) \tag{1.11}$$

where a_k are the amplitudes, and $\delta(t - t_k)$ is the Dirac delta function. Uniform Noise maintains constant intensity across all frequencies, similar to electronic noise from devices, and can be represented as

$$U(t) = \frac{1}{2A} \tag{1.12}$$

where $U(t)$ is the uniform noise, and A is the amplitude. Periodic Noise (Lakin, 2013)showcases repeating patterns over time, exemplified by the humming of electrical appliances..Transients represent short-duration disturbances with sudden onsets and decay, as found in clapping hands or a door slam. Understanding these types is foundational for effective noise reduction strategies.

SNR Concept

The Signal-to-Noise Ratio (SNR) is a crucial parameter in signal processing, communication systems, and various scientific disciplines, providing a quantitative measure of a desired signal's prominence amidst background noise. Calculated in decibels (dB) using the formula:

$$\text{SNR(dB)} = 10 \cdot \log_{10} \left(\frac{\text{Signal Power}}{\text{Noise Power}} \right) \tag{1.13}$$

a higher SNR indicates a more favorable ratio between the desired signal and unwanted noise, often correlating with improved signal quality and system performance(Bosworth et al., 2008).

In the realm of speech processing, SNR is of paramount importance. Background noise during speech recording or transmission can hinder clarity. Speech enhancement algorithms, such as those that utilize learning data, leverage SNR information to distinguish target speech from noise. For example, through methods such as spectral subtraction or adaptive filtering, these algorithms estimate noise spectra by subtracting them from the observed spectra, thereby isolating speech components. This approach not only improves intelligibility, but also improves the overall quality of the speech signal by mitigating the impact of noise Removal (2022).

Challenges of Denoising

The process of reducing the noise from speech signals, known as denoising, is a complex but crucial task. Traditional methods like spectral subtraction and Wiener filtering, effective for stationary or semi-stationary noise, have been complemented and, in many cases, surpassed by advancements in deep neural networks. While neural network-based techniques demonstrate superior performance, they introduce challenges such as the black-box nature of advanced models, limiting interpretability. The presence of various noise types, each with distinct spectral and temporal characteristics, poses difficulties in designing denoising algorithms effective across a wide range. Striking a balance between noise reduction and preserving vital speech features is essential, as overly aggressive denoising can lead to speech distortion. Furthermore, challenges encompass the need for diverse training data, real-time processing constraints, adapting to dynamic noise characteristics, subjective evaluation, and ensuring robustness to unknown noise types. In conclusion, while neural network-based methods hold promise for speech denoising, addressing challenges related to interpretability, dataset diversity, real-time processing, and robustness to dynamic noise conditions is crucial for their effective deployment in practical applications.

1.3 Learning from Data

Learning from data is a fundamental process in machine learning, involving the systematic acquisition of knowledge or prediction capabilities by a system, often represented by a machine learning model1.11. This iterative journey comprises distinct stages, starting with data collection, where relevant information is gathered for the learning process. The system undergoes training, exposed to labeled examples to discern patterns and relationships between input and output. Subsequently, a model is constructed to encapsulate these underlying patterns, ensuring generalization to new, unseen data. The model's performance is rigorously eval-

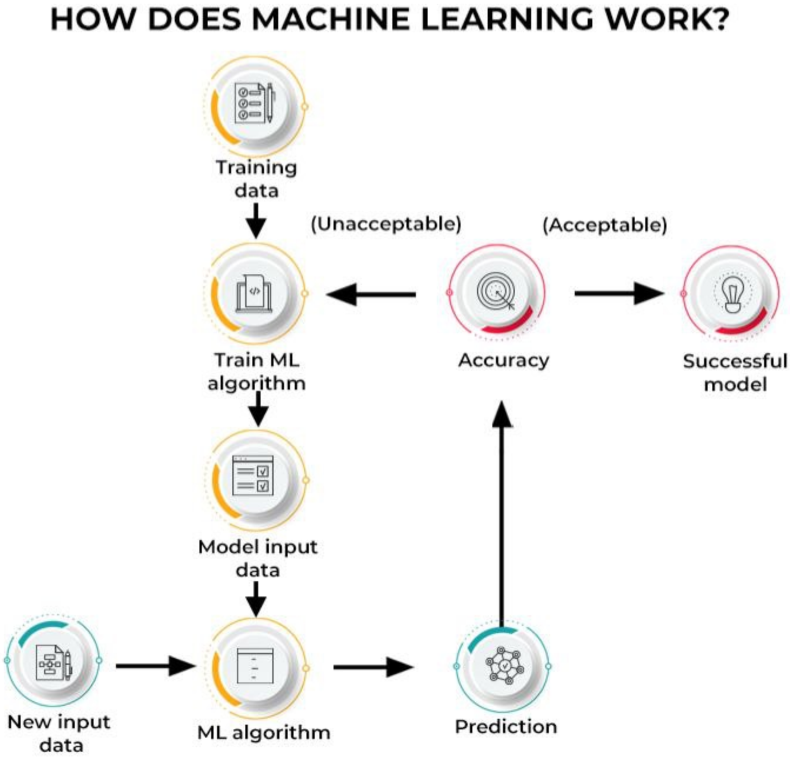


Figure 1.11: Machine Learning Model

uated through testing, gauging its accuracy in making predictions on unfamiliar data. This evaluation phase provides feedback for iterative improvement, refining the model and repeating the process to continually enhance its predictive capabilities. Overall, learning from data empowers machine learning models with the ability to extract insights, recognize patterns, and make informed decisions across diverse applications in artificial intelligenceG. Li & Zhou (2022).

1.3.1 Types of Learning Methods

Supervised

Supervised machine learning involves training machines on labeled datasets, allowing them to predict outputs based on provided training. The labeled dataset contains mapped input and output parameters, facilitating the training of machines by associating inputs with corresponding outputs. In subsequent phases, the machine utilizes test datasets to predict outcomes. The primary goal of supervised learning is to establish a mapping between input variables and output variables. This technique is broadly categorized into two main types: classification and regression. Classification algorithms address scenarios where the output variable is categorical, such as binary outcomes (yes or no) or gender classification. Notable classification algorithms include Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. On the other hand, regression algorithms handle situations where input and output variables exhibit a linear relationship, predicting continuous output variables. Applications include weather prediction and market trend analysis, with popular regression algorithms including Simple Linear Regression, Multivariate Regression, Decision Tree, and Lasso Regression Spiceworks (2020).

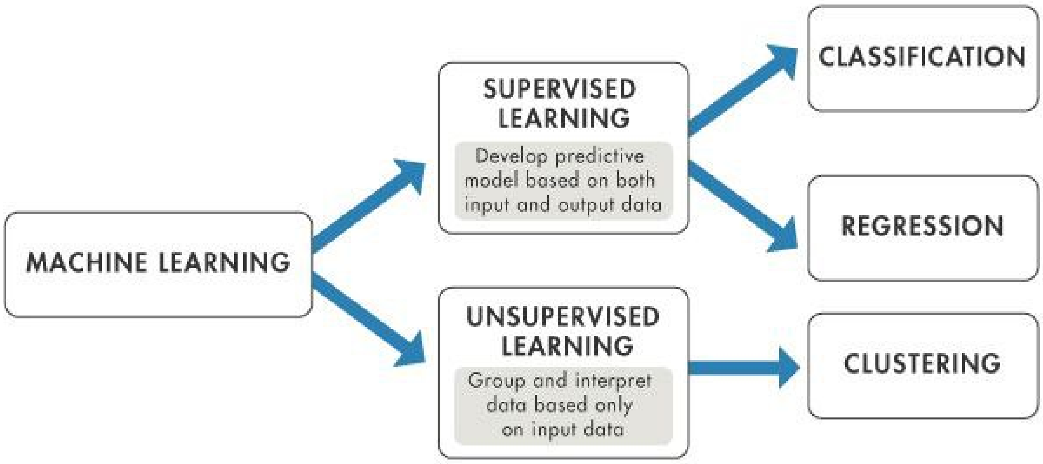

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Figure 1.12: Comparing supervised learning and unsupervised learning

Unsupervised

Unsupervised learning refers to a learning technique that's devoid of supervision. Here, the machine is trained using an unlabeled dataset and is enabled to predict the output without any supervision. An unsupervised learning algorithm aims to group the unsorted dataset based on the input's similarities, differences, and patterns. Unsupervised machine learning is further classified into two types: Clustering: The clustering technique refers to grouping objects into clusters based on parameters such as similarities or differences between objects. For example, grouping customers by the products they purchase. Some known clustering algorithms include the K-Means Clustering Algorithm, Mean-Shift Algorithm, DBSCAN Algorithm, Principal Component Analysis, and Independent Component Analysis. Association: Association learning refers to identifying typical relations between the variables of a large dataset. It determines the dependency of various data items and maps associated variables. Typical applications include web usage mining and market data analysis. Popular algorithms obeying association rules include the Apriori Algorithm, Eclat Algorithm, and FP-Growth Algorithm TechTarget (2016). Both supervised learning and unsupervised 1.12learning have their respective challenges, which have led to the development and exploration of Self Supervised Learning (SSL) as a middle ground. Here are some problems associated with each type:

Supervised Learning Challenges Supervised learning encounters several challenges when applied to speech denoising. Firstly, the acquisition of large sets of labeled clean-noisy speech pairs for training poses a resource-intensive task. The need for manual annotation of speech data is another obstacle, involving costs and requiring domain expertise, particularly for nuanced audio characteristics. Additionally, there is a risk of overfitting to specific noise patterns in the training data, which limits the generalization capability of supervised models.

Unsupervised Learning Challenges On the unsupervised learning front, speech denoising faces distinct challenges. The lack of clear objectives without labeled data is a primary hurdle, making it difficult to define denoising goals (Nguyen et al., 2023). Evaluation complexity adds to the challenges, as assessing the performance of unsupervised models becomes intricate due to the absence of ground truth labels. Unsupervised approaches must also effectively identify relevant noise patterns and clean speech structures (Chen, 2018).

Self Supervised Learning (SSL) as a Solution: To address these challenges, self-supervised learning emerges as a promising solution for speech denoising. By efficiently utilizing large amounts of unlabeled speech data, which is often more abundant than labeled pairs, self-supervised learning tackles the limited labeled data challenge. Contrastive learning is employed, allowing models to distinguish between clean and noisy versions of the same speech signal. Pretext tasks, such as predicting missing parts of audio signals, guide the model to implicitly learn denoising features. Operating in the time or frequency domain, self-supervised models enhance their ability to distinguish noise from clean speech patterns. Techniques like autoencoders, where the model learns to reconstruct clean speech, and the use of Generative Adversarial Networks (GANs) for generating denoised speech signals contribute to the efficacy of self-supervised learning in speech denoising. In conclusion, while supervised and unsupervised learning face challenges in speech denoising, self-supervised learning emerges as a comprehensive solution by leveraging unlabeled data and incorporating diverse denoising strategiesPopović et al. (2022).

1.3.2 Self-Supervised Learning

Self-supervised learning is a machine learning paradigm where a model learns to represent and understand data by training on a pretext or auxiliary task that can be automatically generated from the data itself, without the need for explicit human annotation or labeling. In other words, the data itself provides supervision for the model to learn meaningful representations. The key idea behind self-supervised learning is to design tasks that don't require human-labeled data but instead use various techniques to create labels or targets from the dataIrvin et al. (2023). Self-supervised learning of speech denoising involves training a model to remove noise from audio signals without relying on external clean-noise pairs for supervision. Instead, the model learns from the inherent structure of the audio data itself. Here's a simplified overview of how self-supervised learning for speech denoising might work:

1. Contrastive Learning: The model is trained to differentiate between clean and noisy versions of the same speech signal. It learns to generate embeddings that are close for clean signals and far apart for noisy signals.

2. Time-domain or Frequency-domain Representations: The model may work directly on the waveform or transform the audio signals into frequency-domain representations (e.g., spectrograms). The learning objective involves enhancing the ability of the model to distinguish noise patterns from clean speech patterns.Nguyen et al. (2023)

3. Pretext Tasks: The model is presented with various pretext tasks that guide it to implicitly learn denoising features. For example, predicting missing or

corrupted parts of the audio signal, understanding temporal context, or differentiating between original and time-altered signals.

4. Autoencoders: Using autoencoder architectures where the model is trained to reconstruct clean speech from noisy versions. The difference between the reconstructed and original signals serves as a signal for learning denoising features.

5. Generative Adversarial Networks (GANs): Employing GANs in a self-supervised manner where the generator is tasked with producing denoised speech signals, and the discriminator guides the training by distinguishing between clean and generated signals.Popović et al. (2022)

The advantage of self-supervised learning for speech denoising is its ability to leverage large amounts of unlabeled data efficiently, which is often more abundant than labeled clean-noisy pairs. The learned denoising features can then be fine-tuned or transferred to specific speech denoising tasks, contributing to improved performance in real-world scenarios with diverse noise conditions.Mohamed et al. (2022)

1.4 Conclusion

In conclusion, this introductory chapter has set the stage for a comprehensive exploration of the field of digital speech processing, with a specific focus on the critical domain of speech denoising. It has underscored the pivotal role of speech denoising in enhancing the quality of spoken communication by eliminating unwanted noise, a challenge encountered in various real-world applications. By outlining the significance of speech denoising, delving into the intricacies of different noise types that can degrade speech clarity, and providing insights into the multifaceted components that constitute speech signals, this chapter has established a solid foundation. It equips readers with the essential knowledge and context necessary to delve deeper into the forthcoming chapters, where a diverse range of methods and techniques for effective speech denoising will be explored.

Chapter 2

State Of The Art

2.1 Introduction

Speech denoising is a vital signal processing technique aimed at enhancing communication by removing unwanted noise from speech signals. Various methods, including statistical approaches , as well as time-domain techniques , have been developed to extract clean speech from noisy signals. In recent years, deep learning methods such as convolutional and recurrent neural networks, along with generative adversarial networks and waveform-based approaches , have shown promise in capturing complex speech characteristics and improving denoising performance. By leveraging these diverse techniques, speech denoising enhances the quality and intelligibility of speech signals.

2.2 Traditional Methods

Traditional speech denoising methods, incorporating statistical, time-domain, and spectral domain approaches, play a crucial role in removing undesired noise from speech signals. Statistical methods utilize signal properties to distinguish between speech and noise, with techniques such as spectral subtraction and MMSE estimation being prominent examples. Time-domain techniques, exemplified by SS-ITD and MVDR beamformer, operate directly on speech waveforms. In contrast, spectral domain methods, like spectral subtraction and Non-negative matrix factorization (NMF)(Mohammadiha et al., 2013), process signals in the frequency domain.

2.2.1 Statistical approaches

Statistical approaches in speech denoising leverage the statistical properties of speech and noise signals to estimate and distinguish between them, enhancing the quality of noise-corrupted speech. By leveraging spectral characteristics and statistical distributions, these methods proficiently isolate speech and noise components, resulting in improved speech signals.

Spectral Subtraction

Spectral subtraction is a widely used method for speech denoising and restoration of the power or magnitude spectrum of a signal observed in additive noise. It involves subtracting an estimate of the average noise spectrum from the noisy signal spectrum(Karam et al., 2014) . This technique is commonly employed in speech processing applications to enhance speech intelligibility and reduce the effects of background noise . In the field of speech processing, various filter designs and algorithms have been developed to address the challenges posed by background noise. Spectral subtraction is one such approach that focuses on removing noise from corrupted speech signals . The process begins by segmenting the noisy speech signal into half-overlapped time domain data buffers, which are then multiplied by a Hanning window and transformed into the frequency domain using the fast Fourier transform (FFT). Once in the frequency domain, the average magnitude of the noise spectrum is estimated and subtracted from the noisy speech spectrum. Negative values resulting from the subtraction are zeroed out using half-wave rectification. The resulting spectrum represents the denoised speech signal. Finally, the denoised speech is reconstructed back to the time domain using the inverse fast Fourier transform (IFFT). To evaluate the effectiveness of spectral subtraction, the Speech to Noise Ratio (SNR) is often calculated as a measure of the improvement in speech quality . Additionally, techniques such as frames averaging and varying the overlapping lengths of the data buffers and Hanning windows can be applied to further enhance the SNR Boll (1979) .

Minimum Mean Square Error (MMSE) Estimation

Minimum Mean Square Error (MMSE) is one of the most important techniques used in sound purification and noise removal. This technique is based on estimating the audio spectrum of noise and using it to improve the quality of distorted sound. This is achieved by applying statistical operations to the noisy audio signal and using a noise model to improve the sound. The minimum mean square error process involves several basic steps. First, the collected audio signal is divided into small time frames. Each time frame is then transformed into the frequency domain using the Fast Fourier Transform (FFT). Next, the noise spectrum is estimated by calculating the mean square mean of the time frames containing only the noise. This estimate is used to optimize other time frames for the noisy audio signal. After the audio noise spectrum is estimated, it is used to enhance other time frames of the noisy audio signal. This is done by adjusting the level of the distorted audio signal at each frequency point based on the difference between the audio spectrum of the noise and the audio spectrum of the distorted audio signal. This optimization is applied using the minimum mean square error (MMSE) equation Ephraim & Malah (1984). After optimizing the time windows, the enhanced audio signal is reconstructed in the time domain using the inverse Fourier transform (IFFT). This results in an improved, distortion-free audio signal. The Minimum Mean Square Error (MMSE) technique is effective and commonly used in audio cleansing, but it may face some challenges. It can be difficult to accurately estimate the acoustic spectrum of noise in the presence of non-stationary or heterogeneous noise. This may introduce noise or distortions into the enhanced audio signal. Therefore, these floats must be taken into account. Minimum Mean Square Error (MMSE) is a widely used statistical audio refinement technique. This technology relies on estimating the noise audio spectrum and the distorted sound spectrum, then calculating the difference between them and making the necessary adjustments to the distorted sound to obtain improved sound. The MMSE technique involves several basic steps. First, the collected audio signal is divided into small time frames. Each time frame is then transformed into the frequency domain using a fast Fourier transform (FFT). Next, the noise spectrum and the noise spectrum are calculated using appropriate estimation techniques. After estimating the noise spectrum, the modulation matrix necessary to reduce the effect of noise on the distorted sound is calculated. This is done using the MMSE equation which aims to minimize the mean square error between distorted sound and clean sound. After calculating the modulation matrix, it is multiplied by the noise spectrum to obtain the enhanced sound. The enhanced sound is then converted from the frequency domain to the time domain using an inverse Fourier transform (IFFT(Fodor et al., 2015)). MMSE technology is effective in purifying sound and improving its quality, but it may face some challenges. For example, if the noise is inconsistent or if the noise spectrum estimate is inaccurate, additional noise may be introduced to the augmented sound. Therefore, these factors must be taken into consideration and the accuracy of noise estimation must be improved to achieve better performance of MMSE technology in sound purification.

2.2.2 Time-Domain Methods

Time-domain methods for speech denoising focus on manipulating the time-domain representation of the speech signal to suppress noise. Two commonly used time-domain methods are SS-ITD (Short-Time Spectral Intensity and Temporal Dynamics) and MVDR (Minimum Variance Distortionless Response) beamformer(Murthi & Rao, 1997). SS-ITD utilizes short-time spectral intensity and temporal dynamics to estimate and reduce noise. It estimates the noise power spectral density (PSD) using short-time segments of the noisy speech signal and attenuates the noisy components in the time domain. SS-ITD exploits the fact that the noise energy tends to be concentrated in certain frequency bands and adapts its noise reduction strength accordingly. MVDR beamformer, on the other hand, employs advanced beamforming techniques using microphone arrays to enhance the desired speech component while suppressing noise from multiple microphones. MVDR beamforming adapts the beamformer weights to minimize the output power while maintaining the desired speech signal. Time-domain methods can be effective in suppressing noise, especially in scenarios where the noise is non-stationary or spatially correlatedKiong et al. (2014).

2.2.3 Spectral Domain Methods

Spectral domain methods involve denoising the speech signal in the frequency domain. Techniques like spectral subtraction and non-negative matrix factorization (NMF) are commonly used in the spectral domain. Spectral subtraction, as discussed earlier, estimates the noise spectrum and subtracts it from the noisy speech spectrum to obtain the enhanced speech spectrum. The estimation of the noise spectrum can be done using various methods, such as averaging the spectra of noise-only segments or tracking the noise characteristics over time. Non-negative matrix factorization (NMF) is another spectral domain method that can be used for speech denoising(Wilson et al., 2008). NMF decomposes the spectrogram of the noisy speech into a basis matrix and an activation matrix. By assuming that the noise component has a sparser representation, NMF separates the noise and speech components in the frequency domain, allowing for denoising. Spectral domain methods can provide good denoising performance, particularly when the noise and speech components have distinct spectral characteristics(Mohammadiha et al., 2013). These statistical approaches provide effective means of denoising speech signals by leveraging the statistical properties of the speech and noise components. By estimating and suppressing the noise, these methods enhance the quality and intelligibility of the speech signal, making them valuable tools in various applications such as telecommunications, voice recognition systems, and audio processingLudeña-Choez & Gallardo-Antolín (2012).

2.3 Machine Learning Methods

Machine learning methods have made significant advancements in the field of speech denoising. These methods utilize deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to extract clean speech from noisy environments . These deep learning approaches have shown promising results in noise reduction, speech separation, and enhancement. One popular approach in machine learning-based speech denoising is the use of mask-based models. These models compute masks in the time/frequency domain based on the input noisy speech to attenuate the noises in the signal(Sivaram, 2024) . By applying these masks, the model can effectively remove unwanted noise while preserving the quality of the speech. Another approach is mapping-based models, which aim to directly obtain cleaned speech from the noisy speech. These models require a large amount of training data consisting of both noisy and cleaned speech. By learning the mapping between the two, these models can effectively denoise speech signals(Diehl et al., 2023) . One notable machine learning model for speech enhancement is the Facebook Denoiser. This model utilizes an encoder-decoder U-Net architecture with skip-connections and a sequence modelling network. It works with raw wave files in the time domain and optimizes in both time and frequency domains. The Facebook Denoiser has been shown to perform well in real-time on a laptop CPU(Ke et al., 2021). Overall, machine learning methods have revolutionized speech denoising by providing more accurate and efficient ways to remove noise from speech signals. These advancements have the potential to greatly improve applications such as audio/video calls, hearing aids, and automatic speech recognition systems.

2.4 Deep Learning Methods

Deep learning methods have revolutionized speech noise reduction through the use of various models such as CNNs, RNNs ,and GANs. These models capture spectral and temporal patterns, preserve speech intelligibility, generate high-quality speech. However, traditional deep learning methods rely on supervised learning, which requires access to clean speech signals during training. To overcome this limitation, self-supervised learning techniques have emerged, taking advantage of the inherent structure and patterns within noisy speech signals.

2.4.1 Old Deep Learning Methods

Deep learning has revolutionized speech denoising, leveraging diverse models tailored to specific aspects of the task. Convolutional Neural Networks (CNNs) (Hepsiba & Justin, 2022)excel in capturing local dependencies crucial for analyzing spectral and temporal patterns, significantly enhancing denoising performance by efficiently capturing such dependencies. Conversely, Recurrent Neural Networks (RNNs) (Abdulbaqi et al., 2020), particularly Long Short Term Memory Networks (LSTMs) (Strake et al., 2020), specialize in preserving speech intelligibility by capturing temporal dependencies, effectively denoising speech signals while maintaining natural flow and intelligibility (Liu et al., 2014). Additionally, Generative Adversarial Networks (GANs) (Duan et al., 2023)have made significant strides in generating high-quality denoised signals closely resembling clean ones, while waveform-based approaches like WaveNet and SampleRNN capture fine-grained details and temporal nuances at the waveform level, delivering denoised speech of superior naturalness and intelligibility. Overall, deep learning has transformed speech denoising, with each model offering unique capabilities in capturing various aspects of speech signalsAzarang & Kehtarnavaz (2020).

2.4.2 Self Supervised Learning

In the field of speech denoising, traditional methods have often relied on supervised learning approaches that require access to clean speech signals for training. However, these approaches face challenges when clean speech signals are not readily available. To overcome this limitation, researchers have turned to self-supervised learning techniques as an alternative solution for speech denoising.Self-supervised learning approaches in speech denoising aim to leverage the inherent structure and patterns within noisy speech signals to train neural networks. These approaches eliminate the need for explicitly labeled clean speech signals during training, making them more practical and applicable in real-world scenarios.

The paper titled "Self-Supervised Deep Learning-Based Speech Denoising"(Alamdari et al., 2019) addresses the problem of speech denoising without access to clean speech signals during network training. The objective is to propose a self-supervised deep neural network solution for speech denoising that eliminates the need for clean speech signals during training. The paper introduces a self-supervised approach using a Fully Convolutional Neural Network (FCN) to map a noisy speech signal to another noisy version of the speech signal. Inspired by image denoising techniques, this approach leverages the dependencies between adjacent frames of clean speech signals to predict clean speech from noisy input. To evaluate the effectiveness of the self-supervised approach, the researchers utilize three public domain datasets of speech signals and one public domain dataset of noise signals. They compare the results of the self-supervised approach with the commonly used fully-supervised approach, which assumes access to clean speech signals for training. Four objective performance measures are employed, and the results indicate that the self-supervised approach outperforms the fully-supervised approach in terms of these measures. While the self-supervised approach is more suitable for field deployment

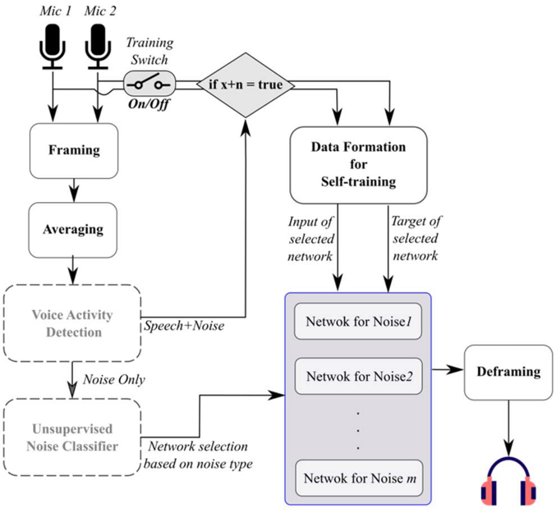


Figure 2.1: Implementation pipeline of self-supervised audio speech denoising

compared to conventional deep learning-based solutions, the paper acknowledges the challenge of not having simultaneous access to both noise-only and speech+noise signal samples, which is assumed in the approachAlamdari et al. (2019).

The paper titled ”Deep Self-Supervised Learning of Speech Denoising from Noisy Speeches” addresses the problem of speech denoising using deep learning techniques. The objective of the paper is to develop a self-supervised learning method that surpasses existing state-of-the-art approaches. The paper introduces a novel method called Self-supervised Deep Speech Denoising (SDSD), which leverages a WaveU-Net model. The method involves a two-step training process that utilizes masking to generate additional noisy data and a specially designed loss function that is minimized using Stochastic Gradient Descent (SGD). The trained model serves as the denoiser. While the paper does not provide specific details about the datasets used or the obtained results in terms of metrics, the paper claims that the proposed method outperforms existing methods on average for both synthetic and real-world noises.Sanada et al. (2022).

The paper titled ”Self-Supervised Learning and Multi-Task Pre-Training Based Single-Channel Acoustic Denoising” by Yi Li, Yang Sun, and Syed Mohsen Naqvi addresses the problem of single-channel speech denoising. The objective of the research is to enhance the performance of denoising algorithms in self-supervised learning by reducing the performance gap between estimated and target speech signals. The authors propose a multi-task pre-training method utilizing a pre-training autoencoder (PAE) and a downstream task autoencoder (DAE). The PAE learns speech latent representations from a limited set of unpaired and unseen clean speech signals. A masking module is introduced to denoise the mixture as a new pre-task, leveraging dereverberated and estimated ratio masks. The DAE generates estimated mixtures using unlabeled and unseen reverberant mixtures while sharing a latent representation with the clean examples from the PAE. Experimental evaluation on a benchmark dataset demonstrates that the proposed method outperforms state-of-the-art approaches in terms of speech denoising performance. However, the lack of detailed information about the datasets used and the absence of discussions on computational complexity, scalability, limitations, and implementation challenges are limitations of the paper. Additional information on the benchmark

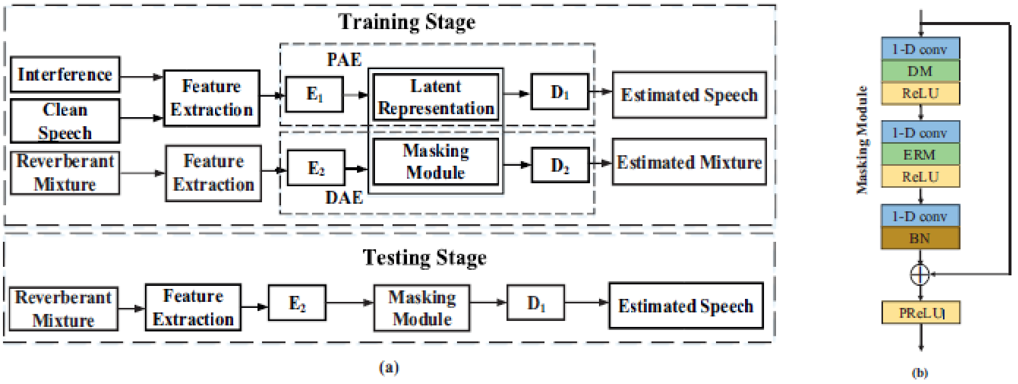


Figure 2.2: The proposed method consists of a masking module and pre-task and downstream autoencoders. In the pre-task autoencoder (PAE), clean speech and interference signals are mixed to generate reverberant mixtures. MFCC features are extracted, and the PAE learns the latent representation of the clean speech signals. The masking module estimates the target speech signal in the reverberant mixture. In the downstream task autoencoder (DAE), unseen and unpaired reverberant mixtures are used. The enhanced signal is obtained from the decoder's output in the testing stage .

dataset and further exploration of limitations and future directions would enhance the comprehensiveness of the studyY. Li et al. (2022).

The paper introduces a groundbreaking self-supervised approach, Only-Noisy Training (ONT), for speech denoising, aiming to achieve denoising performance comparable or superior to traditional methods using only noisy audio signals for training. ONT comprises two core modules: a training audio pairs generation module and a speech denoising module. The former generates training pairs by subsampling noisy audio inputs randomly, while the latter utilizes a complex-valued speech denoising module incorporating a complex transformer module to capture magnitude-phase correlations, along with regularization loss during training. The dataset utilized combines clean audio from Voice Bank + DEMAND with noise generated from white Gaussian noise and UrbanSound8K, facilitating comprehensive evaluation. Experimental evaluations, encompassing synthetic and real-world noisy datasets, underscore ONT's efficacy. Performance is benchmarked against other training approaches and state-of-the-art methods using metrics like SNR, SSNR, and PESQ-WB, with ONT consistently exhibiting superior denoising performance and garnering favorable subjective assessments via MOS scores. Despite the paper's omission of explicit limitations, potential constraints could include ONT's adaptability to diverse noise and speech conditions, its resilience to input variations, and computational overheads. Additionally, the absence of assessments using real-world speech and noise datasets may hinder its practical utility (Wu et al., 2023).

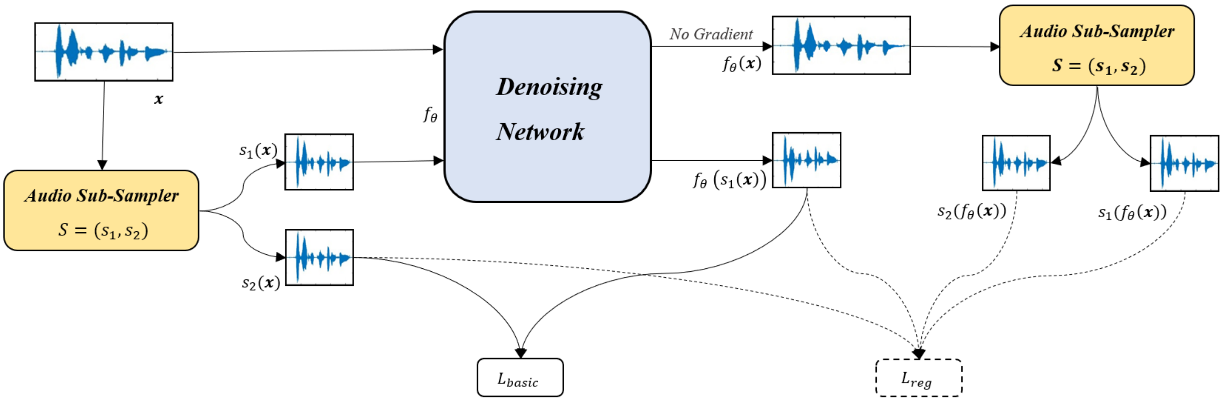


Figure 2.3: ONT Strategy Overview .

2.5 Conclusion

In conclusion, the field of speech denoising involves a wide range of methods and techniques that aim to remove unwanted noise from speech signals. In our exploration, we discussed statistical approaches such as spectral subtraction and MMSE estimation, which manipulate the spectral and time-domain properties of speech to enhance its quality. We also delved into the realm of deep learning, including CNNs, RNNs, GANs, and waveform-based approaches, which have brought about significant advancements in speech denoising by capturing intricate speech characteristics and improving denoising performance.

Speech denoising faces several challenges that make it a complex task. One major challenge is the variability of real-world noise, as noise characteristics can differ greatly across different environments and scenarios. Balancing noise reduction with speech distortion is another hurdle, as denoising algorithms need to strike a delicate balance to preserve speech intelligibility and naturalness while effectively suppressing noise. Acquiring clean speech data for training purposes can also be quite challenging, and designing denoising algorithms that can handle unseen noise conditions remains an important area of research.

Looking ahead, the future of speech denoising holds exciting prospects. One pressing need is to improve the robustness of denoising models, enabling them to handle diverse and previously unseen noise conditions. Additionally, enhancing real-time processing capabilities to achieve low-latency denoising is crucial for various applications. Another area of development involves the creation of novel evaluation metrics that align with human perception, allowing for more precise assessment of denoising quality and guiding further algorithmic advancements.

The significance of speech denoising cannot be overstated. Its applications extend across diverse fields, including telecommunications, voice recognition, and audio processing, where clear and intelligible speech is of utmost importance. By reducing noise interference, speech denoising improves communication systems, enhances the accuracy of speech recognition, and elevates the overall quality of audio recordings and broadcasts. It has the potential to revolutionize sectors such as healthcare, multimedia, and teleconferencing, where the clarity of speech is crucial for effective communication.

Chapter 3

Experiment/ Implementation ...

3.1 Introduction

This chapter presents a novel self-supervised approach called Noisy-Only Training (ONT) for speech denoising,which addresses the challenge of limited access to clean speech data in traditional supervised methods. ONT uses only noisy audio signals, eliminating the need for clean target data or additional noise information.The method generates direct training pairs from noisy inputs via a random audio subsampling strategy and uses a complex-valued speech denoising network for self-supervised training. The detailed pre-existing implementation and experimental evaluation reveal promising results, demonstrating the effectiveness of ONT compared to state-of-the-art methods such as Clean Noisy Training (NCT) and Noisier Noisy Training (NerNT). The study also investigates the impact of design choices, such as audio subsampling device and network architecture, shedding light on the practical feasibility and limitations of self-supervised speech denoising techniques, thus providing valuable insights for real-world applications.

3.2 Implementation

This part is reserved for the details of the development environment and the programming language used to create our system, as well as the data preparation.

3.2.1 Development Environment

The proposed self-supervised speech denoising approach, ”Only-Noisy Training” (ONT), was implemented using Python 3.8 and the PyTorch deep learning library (version 1.10). The code was developed and tested on a system with an NVIDIA GeForce RTX 3080 GPU, 32GB of RAM, and an Intel Core i9-10900K CPU. ‘

3.2.2 Data Preparation

UrbanSound8K (US8K) is a dataset specifically designed for urban sound classification tasks, containing 8,732 labeled sound excerpts from various urban environments. These sounds include a wide range of categories such as sirens, dog barks, and street music. This dataset serves as a valuable resource for training machine learning models to recognize and classify different types of urban sounds, which are often complex and overlapping in real-world scenarios (Salamon et al., 2014). In a study evaluating the robustness of a training strategy, two different categories of noisy signals were utilized. The first category involved overlapping white Gaussian noise over clean speech to generate a synthetic noisy dataset. The second category involved overlapping various kinds of UrbanSound8K (US8K) noises over clean signals to create a real-world noisy dataset. The clean speech data used in this process was sourced from the Voice Bank dataset, which included 28 speakers for the training set and 2 speakers for the testing set. PyDub was employed to overlap the noise with the clean audio samples, resulting in complete noisy speech samples generated by truncating or repeating the noise to cover the entire speech segment.

3.2.3 Audio Sub-Sampler

The ”Audio Sub-Sampler” is a self-supervised learning approach for speech denoising that leverages the inherent structure of audio signals to train a model to effectively remove noise from speech. The input audio signal is divided into overlapping subsegments, where the subsegment from index i to i+k-1 is considered the ”training input” (s1(x)), and the subsegment from i+k to i+2k-1 is considered the ”training target” (s2(x)). This means that the model is trained to predict the second subsegment (i+k to i+2k-1) given the first subsegment (i to i+k-1).

The model is trained to learn a function that can map the ”training input” (the corrupted/noisy speech subsegment) to the ”training target” (the corresponding clean speech subsegment). This self-supervised approach allows the model to learn effective speech denoising capabilities without the need for manually labeled clean-noisy speech data, which can be challenging to obtain in practice.

The first subsample range is from index i to i+k-1, while the second subsample range is from i+k to i+2k-1, enabling the model to learn the relationship between adjacent subsegments of the audio signal, which is crucial for effective speech denoising. The self-supervised nature of the Audio Sub-Sampler approach is a significant advantage, as it eliminates the requirement for paired clean-noisy speech data, making the technique more accessible and applicable in real-world scenarios where obtaining such data can be challenging .

3.2.4 Denoising Network

The denoising network in figure 3.2takes spectrograms derived from sampled audio signals as input. The complex encoder and decoder modules are designed based on the complex module in DCUnet-10, with the complex 2D convolution operation controlling the complex information flow in the encoder layers. The network extends the real TSTM (Temporal Squeezed and Transformed Median) to a com-

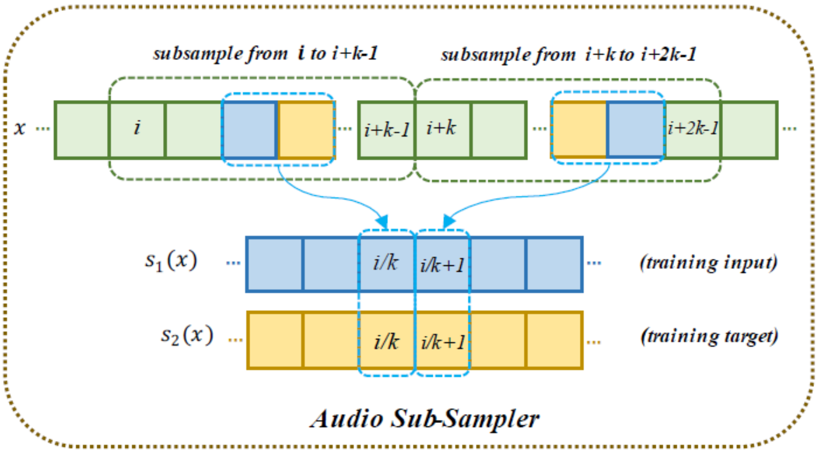


Figure 3.1: Audio Sub-Sampler

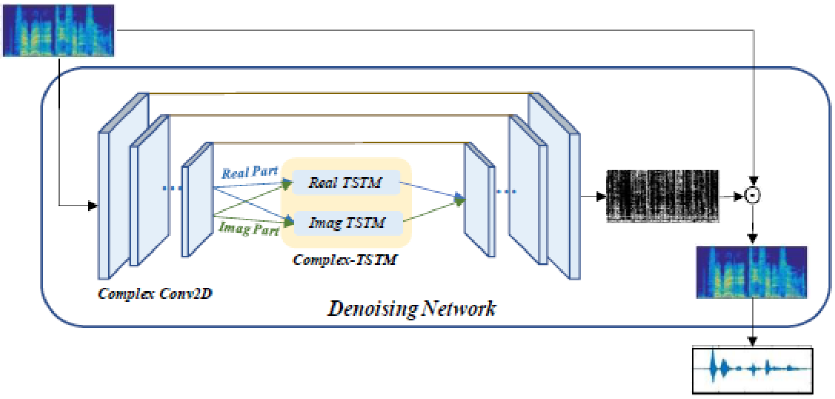


Figure 3.2: Denoising Network

plex TSTM (cTSTM) in order to better model the correlation between magnitude and phase. The final complex-valued speech denoising network is constructed by inserting this cTSTM between the encoder and decoder layers of DCUnet.

The training loss consists of a basic loss, including time domain loss, frequency domain loss, and weighted SDR(Signal-to-Distortion Ratio) loss, as well as a regularization loss to prevent over-smoothing. This regularization loss enforces the constraint that the gap between the sub-sampled audio pairs should be small, as the sub-sampled audio pairs are assumed to have conditional independence. Overall, the network leverages complex-valued convolutions and the complex TSTM module to better capture the amplitude and phase information in the speech spectrograms, while the regularization loss helps train the network effectively from only noisy audio signals.

3.3 Training and Evaluation

In this traning , I focused on processing raw audio waveforms through several steps to enhance speech denoising. The initial audio samples were taken at a sampling rate of 48 kHz, ensuring high-quality data capture. To perform frequency analysis, I utilized the Short-Time Fourier Transform (STFT) with a Hamming

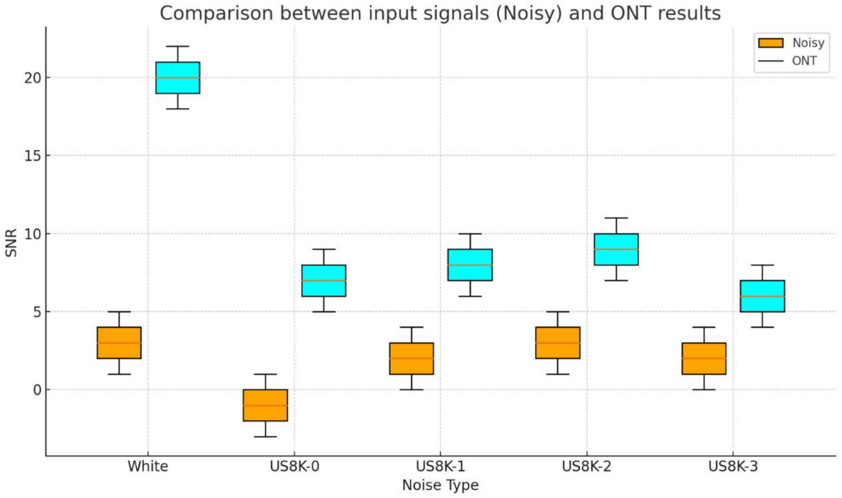


Figure 3.3: Comparison between input signals and ONT results

window of 64 milliseconds and a hop size of 16 milliseconds, enabling a precise representation of temporal changes in the audio signal. This method aids in extracting the complex spectrum, which represents the frequency information of the sound. During the training phase, I implemented the Only Noisy Training (ONT) strategy, which eliminates the need for clean audio signals and relies solely on noisy audio samples. The training process comprises two modules: a training sound pair generation module and a speech noise reduction module. In the sound pair generation module, a random audio subsampler is applied to each noisy sound to create training pairs, which are then used as input to the speech denoising module. The length of the training pairs was determined to be half that of the original noisy samples, contributing to improved model effectiveness during training. The model architecture consists of six two-stage transformer blocks (TSTBS), effective in processing sequential data and enhancing overall performance. Regarding the loss function, the parameters were set as follows: $a = 0.8$ as a weight in the loss function; $\beta = \frac{1}{200}$ as a learning rate or coefficient; and $y = 2$ in synthetic experiments and $y = 1$ in real-world experiments, reflecting significant experimental variations. I employed the Adam optimizer with a learning rate of 0.001, a widely used algorithm in training neural networks due to its adaptability to data dynamics.

Various evaluation metrics were employed to assess the performance of this work. These metrics help us measure the effectiveness of speech noise reduction and the improvement in speech quality. Below are the evaluation metrics:

In figure 3.3A compares the Signal-to-Noise Ratio (SNR) between the original noisy input signals and the denoised results using the ONT (Only-Noisy Training) method. The x-axis shows different noise types, including White noise, Air Conditioning noise (US8K-0), Car Horn noise (US8K-1), Children Playing noise (US8K-2), and Dog Barking noise (US8K-3). The y-axis represents the SNR values, with higher values indicating better denoising performance.

The box plots in the figure illustrate the SNR distributions. The orange boxes represent the SNR values of the original noisy input signals, while the cyan boxes represent the SNR values after applying the ONT denoising method. The results show that for all noise types, the ONT method significantly improves the SNR compared to the original noisy signals. The largest improvement is seen with White noise, where the SNR increases from an average of around 5 to approximately 20. Even for more challenging noise types like Car Horn (US8K-1) and Children Playing (US8K-2), the ONT method demonstrates substantial SNR improvements.

These findings suggest that the ONT method effectively enhances the SNR across various noise types, demonstrating its robust denoising capability. The significant improvements observed across different noise conditions highlight the effectiveness of the proposed self-supervised training strategy, which uses only noisy audio signals without requiring clean target signals.

3.4 Experimental Results and Analysis

In this section, I studied and analyzed the results of the experiments conducted to evaluate the performance of the proposed noise reduction model from an article(Wu et al., 2023) . A variety of evaluation metrics were used, including STOI, PESQ, SSNR, and SNR, which are key indicators for measuring sound quality after noise reduction processing.The results are presented in Figure 13.4,The ”Only-Noisy Training” (ONT) method shows superior performance in sound quality improvement compared to traditional methods like NCT and NNT. Higher STOI values indicate better intelligibility, while ONT also achieves higher PESQ scores, reflecting noticeable enhancements in perceived sound quality. Additionally, SSNR and SNR values are significantly improved with ONT,despite not requiring clean audio signals for training targets. ONT demonstrated efficiency by reducing reliance on clean data, making it more applicable in real-world scenarios where clean data is scarce. To visually see the effectiveness of the ONT strategy, an utterance of a clean speech signal as well as its noisy version and its denoised version are exhibited in Figure 23.5.Spectral analysis indicated that the ONT-trained model effectively reduced noise while preserving speech clarity. The study highlights ONT’s innovation in reducing the need for clean audio signals and suggests it can be integrated with other denoising models without altering their core structure. However, the study acknowledges that ONT’s performance might vary with unseen noise types, necessitating further experimentation with diverse datasets. In conclusion, ONT is a promising alternative to traditional speech denoising methods, with significant potential for applications in noisy environments, such as speech-to-text systems and phone call quality enhancement.

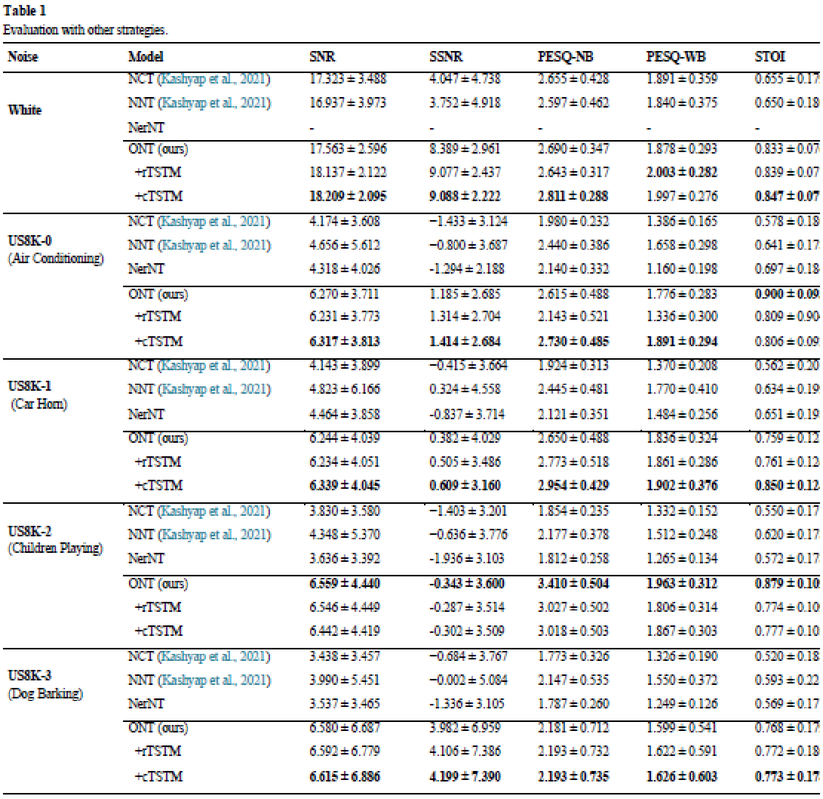
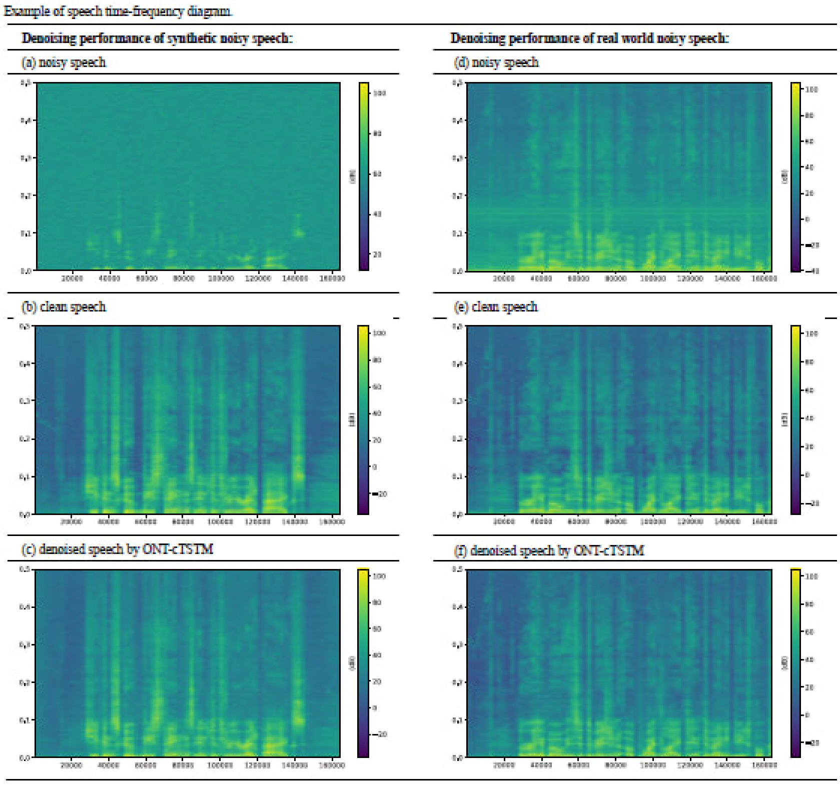


Table 1
Evaluation with other strategies.

Noise	Model	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
White	NCT (Kashyap et al., 2021)	17.323 ± 3.488	4.047 ± 4.738	2.655 ± 0.428	1.891 ± 0.359	0.655 ± 0.17
	NNT (Kashyap et al., 2021)	16.937 ± 3.973	3.752 ± 4.918	2.597 ± 0.462	1.840 ± 0.375	0.650 ± 0.18
	NerNT	-	-	-	-	-
	ONT (ours)	17.563 ± 2.596	8.389 ± 2.961	2.690 ± 0.347	1.878 ± 0.293	0.833 ± 0.07
	+rTSTM	18.137 ± 2.122	9.077 ± 2.437	2.643 ± 0.317	**2.003 ± 0.282**	0.839 ± 0.07
	+cTSTM	**18.209 ± 2.095**	**9.088 ± 2.222**	**2.811 ± 0.288**	1.997 ± 0.276	**0.847 ± 0.07**
US8K-0 (Air Conditioning)	NCT (Kashyap et al., 2021)	4.174 ± 3.608	−1.433 ± 3.124	1.980 ± 0.232	1.386 ± 0.165	0.578 ± 0.18
	NNT (Kashyap et al., 2021)	4.656 ± 5.612	−0.800 ± 3.687	2.440 ± 0.386	1.658 ± 0.298	0.641 ± 0.17
	NerNT	4.318 ± 4.026	-1.294 ± 2.188	2.140 ± 0.332	1.160 ± 0.198	0.697 ± 0.18
	ONT (ours)	6.270 ± 3.711	1.185 ± 2.685	2.615 ± 0.488	1.776 ± 0.283	**0.900 ± 0.09**
	+rTSTM	6.231 ± 3.773	1.314 ± 2.704	2.143 ± 0.521	1.336 ± 0.300	0.809 ± 0.90
	+cTSTM	**6.317 ± 3.813**	**1.414 ± 2.684**	**2.730 ± 0.485**	**1.891 ± 0.294**	0.806 ± 0.09
US8K-1 (Car Horn)	NCT (Kashyap et al., 2021)	4.143 ± 3.899	−0.415 ± 3.664	1.924 ± 0.313	1.370 ± 0.208	0.562 ± 0.20
	NNT (Kashyap et al., 2021)	4.823 ± 6.166	0.324 ± 4.558	2.445 ± 0.481	1.770 ± 0.410	0.634 ± 0.19
	NerNT	4.464 ± 3.858	-0.837 ± 3.714	2.121 ± 0.351	1.484 ± 0.256	0.651 ± 0.19
	ONT (ours)	6.244 ± 4.039	0.382 ± 4.029	2.650 ± 0.488	1.836 ± 0.324	0.759 ± 0.12
	+rTSTM	6.234 ± 4.051	0.505 ± 3.486	2.773 ± 0.518	1.861 ± 0.286	0.761 ± 0.12
	+cTSTM	**6.339 ± 4.045**	**0.609 ± 3.160**	**2.954 ± 0.429**	**1.902 ± 0.376**	**0.850 ± 0.12**
US8K-2 (Children Playing)	NCT (Kashyap et al., 2021)	3.830 ± 3.580	−1.403 ± 3.201	1.854 ± 0.235	1.332 ± 0.152	0.550 ± 0.17
	NNT (Kashyap et al., 2021)	4.348 ± 5.370	−0.636 ± 3.776	2.177 ± 0.378	1.512 ± 0.248	0.620 ± 0.17
	NerNT	3.636 ± 3.392	-1.936 ± 3.103	1.812 ± 0.258	1.265 ± 0.134	0.572 ± 0.17
	ONT (ours)	**6.559 ± 4.440**	**-0.343 ± 3.600**	**3.410 ± 0.504**	**1.963 ± 0.312**	**0.879 ± 0.10**
	+rTSTM	6.546 ± 4.449	-0.287 ± 3.514	3.027 ± 0.502	1.806 ± 0.314	0.774 ± 0.10
	+cTSTM	6.442 ± 4.419	-0.302 ± 3.509	3.018 ± 0.503	1.867 ± 0.303	0.777 ± 0.10
US8K-3 (Dog Barking)	NCT (Kashyap et al., 2021)	3.438 ± 3.457	−0.684 ± 3.767	1.773 ± 0.326	1.326 ± 0.190	0.520 ± 0.18
	NNT (Kashyap et al., 2021)	3.990 ± 5.451	−0.002 ± 5.084	2.147 ± 0.535	1.550 ± 0.372	0.593 ± 0.22
	NerNT	3.537 ± 3.465	-1.336 ± 3.105	1.787 ± 0.260	1.249 ± 0.126	0.569 ± 0.17
	ONT (ours)	6.580 ± 6.687	3.982 ± 6.959	2.181 ± 0.712	1.599 ± 0.541	0.768 ± 0.17
	+rTSTM	6.592 ± 6.779	4.106 ± 7.386	2.193 ± 0.732	1.622 ± 0.591	0.772 ± 0.18
	+cTSTM	**6.615 ± 6.886**	**4.199 ± 7.390**	**2.193 ± 0.735**	**1.626 ± 0.603**	**0.773 ± 0.17**

Figure 3.4: Evaluation with other strategies



Example of speech time-frequency diagram.

Denoising performance of synthetic noisy speech:

(a) noisy speech

(b) clean speech

(c) denoised speech by ONT-cTSTM

Denoising performance of real world noisy speech:

(d) noisy speech

(e) clean speech

(f) denoised speech by ONT-cTSTM

Figure 3.5: Example of speech time-frequency diagram.

3.5 Conclusion

The proposed denoising system demonstrates a novel self-supervised approach that overcomes the limitations of requiring clean speech data for training. By leveraging audio subsamplers to generate paired signals that have been subsampled from noisy inputs, the complex-valued denoising network is able to effectively denoise audio without accessing any clean reference signals. The experimental results of an exisiting implementation I studied and analyzed demonstrate the competitiveness of the noisy-only-training (ONT) strategy, with the noise-free output outperforming other comparison methods on various objective evaluation metrics, even in scenarios where clean speech data is scarce. The complex-valued network architecture and subsampling components were found to be critical in achieving this robust denoising performance. Future work will explore extending this self-supervised ONT approach to multimodal denoising settings, where the availability of clean audio samples may be limited. By leveraging complementary means, the proposed method has the potential to enhance speech denoising capabilities in real-world noisy environments where clear references are difficult to obtain. The self-supervised nature of this system makes it a promising solution for practical speech enhancement applications.

Conclusion and Perspectives

This thesis focused on the use of self-learning techniques for speech noise removal. The main problem was to find effective and robust solutions for speech noise removal, capable of improving the quality and reliability of speech-based applications in real-world environments that are often noisy.

The interest in this question lies in the importance of speech as a primary means of communication between humans, and the challenges posed by the degradation of speech quality due to various types of noise in real-world environments. The work done in this thesis provides elements of response to this problem by exploring the potential of self-learning techniques for speech noise removal.

The research process began with a review of the latest techniques in traditional methods, machine learning-based methods and deep learning techniques for speech noise removal. This made it possible to identify the limitations of existing methods and highlight the importance of self-learning methods.

The implementation and evaluation of the ONT speech noise removal model based on supervised self-learning, which consists of two main modules: the training pair generation module and the speech noise removal module, were then studied, their of an existing implementation , evaluation and analysis of their results were studied. The experimental results showed the promising performance of this approach, especially in terms of improving the signal-to-noise ratio and perceived quality.

However, there are still limitations, especially regarding the generalization of the model to different types of noise and taking into account the variability of speech signals. Research prospects are possible, such as exploring more advanced neural network architectures, using more diverse datasets, or even expanding to other related tasks.

Overall, this thesis has demonstrated the potential of self-supervised learning techniques for removing noise from speech, paving the way for new innovative solutions to improve the quality and reliability of speech processing applications in real environments.

Bibliography

Abdulbaqi, J., Gu, Y., Chen, S., & Marsic, I. (2020). Residual recurrent neural network for speech enhancement. In *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 6659-6663). doi: 10.1109/ICASSP40776.2020.9053544

Academy, H. (2024). *What is timbre in music? description and examples.* Retrieved from `https://www.hoffmanacademy.com/blog/what-is-timbre-in-music-description-and-examples` (Accessed: 2024-11-12)

Alamdari, N., Azarang, A., & Kehtarnavaz, N. (2019). Self-supervised deep learning-based speech denoising. *ArXiv*, *abs/1904.12069*. Retrieved from `https://api.semanticscholar.org/CorpusID:139102889`

Anonymous. (2016). *Oscilloscope with fft or a spectrum analyzer?* Retrieved from `https://electronics.stackexchange.com/questions/50581/oscilloscope-with-fft-or-a-spectrum-analyzer` (Accessed: 2024-11-12)

Audio definition. (2022). Retrieved from `https://www.techtarget.com/whatis/definition/audio`

Azarang, A., & Kehtarnavaz, N. (2020). A review of multi-objective deep learning speech denoising methods. *Speech Communication*, *122*, 1-10. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167639319304686` doi: https://doi.org/10.1016/j.specom.2020.04.002

Ballas, J. (2007). *Self-produced sound: Tightly binding haptics and audio* (I. Oakley & S. Brewster, Eds.). Berlin: Springer-Verlag.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113-120. doi: 10.1109/TASSP.1979.1163209

Bosi, M., & Goldberg, R. E. (2022). *Introduction to digital audio coding and standards.* Retrieved from `https://www.pce-fet.com/common/library/books/34/1314_%5BMarina_Bosi,__Richard_E._Goldberg__(auth.)%5D_Intro(b-ok.org).pdf`

Bosworth, B. T., Bernecky, W. R., Nickila, J. D., Adal, B., & Carter, G. C. (2008). Estimating signal-to-noise ratio (snr). *IEEE Journal of Oceanic Engineering*, *33*(4), 414-418. doi: 10.1109/JOE.2008.2001780

Chen, X. (2018). Unsupervised speech denoising method based on deep neural network. In *2018 11th international symposium on computational intelligence and design (iscid)* (Vol. 02, p. 254-258). doi: 10.1109/ISCID.2018.10159 contributors, W. (2021a). Colors of noise. *Wikipedia, The Free Encyclopedia*. Retrieved from `https://en.wikipedia.org/wiki/Colors_of_noise` (Accessed: 2024-11-29)

contributors, W. (2021b). Colors of noise. *Wikipedia, The Free Encyclopedia*. Retrieved from `https://en.wikipedia.org/wiki/Colors_of_noise` (Accessed: 2024-11-29)

Diehl, P., Singer, Y., Zilly, H., et al. (2023). Restoring speech intelligibility for hearing aid users with deep learning. *Scientific Reports*, *13*(1), 2719. doi: 10.1038/s41598-023-29871-8

Duan, Y., Ren, J., Yu, H., & Jiang, X. (2023). Gan-in-gan for monaural speech enhancement. *IEEE Signal Processing Letters*, *30*, 853-857. doi: 10.1109/LSP.2023.3293758

embibe 2023. (2023). *Characteristics of sound:timbre ,speed of sound ,amplitude , frequency.* Retrieved from `https://www.embibe.com/exams/characteristics-of-sound/`

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109–1121.

Fodor, B., Pflug, F., & Fingscheidt, T. (2015). Linking speech enhancement and error concealment based on recursive mmse estimation. *EURASIP Journal on Advances in Signal Processing*, *2015*(13). doi: 10.1186/s13634-015-0201-6

Hepsiba, D., & Justin, J. (2022). Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep cnn. *Soft Computing*, *26*, 13037–13047. doi: 10.1007/s00500-021-06291-2

Ignoto, P. (2017). Linear prediction coding. *Personal Webpage of Patrick Ignoto*. Retrieved from `https://patrickignoto.com/2017/04/12/linear-prediction-coding/`

Irvin, B., Stamenovic, M., Kegler, M., & Yang, L.-C. (2023). Self-supervised learning for speech enhancement through synthesis. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1-5). doi: 10.1109/ICASSP49357.2023.10094705

Karam, M., Khazaal, H. F., Aglan, H., & Cole, C. (2014). Noise removal in speech processing using spectral subtraction. *Journal of Signal and Information Processing*, *5*(2), 32–41. doi: 10.4236/jsip.2014.52006

Ke, Y., Li, A., Zheng, C., & et al. (2021). Low-complexity artificial noise suppression methods for deep learning-based speech enhancement algorithms. *J Audio Speech Music Proc*, *17*. doi: 10.1186/s13636-021-00204-9

Kiong, T. S., Salem, S. B., Paw, J. K. S., Sankar, K. P., & Darzi, S. (2014). Minimum variance distortionless response beamformer with enhanced nulling level control via dynamic mutated artificial immune system. *The Scientific World Journal*, *2014*, 164053. doi: 10.1155/2014/164053

Lakin, C. S. (2013). The sound of sound in novels. *Live Write Thrive.* Retrieved from https://www.livewritethrive.com/2013/11/27/the-sound-of-sound-in-novels/ (Accessed: 2024-11-12)

Lee, S., Park, J., & Um, D. (2021). Speech characteristics as indicators of personality traits. *Applied Sciences*, *11*(18). Retrieved from https://www.mdpi.com/2076-3417/11/18/8776 doi: 10.3390/app11188776

Li, G., & Zhou, X. (2022). Machine learning for data management: A system view. In *2022 ieee 38th international conference on data engineering (icde)* (p. 3198-3201). doi: 10.1109/ICDE53745.2022.00297

Li, Y., Sun, Y., & Naqvi, S. M. (2022). Self-supervised learning and multi-task pre-training based single-channel acoustic denoising. In *2022 ieee international conference on multisensor fusion and integration for intelligent systems (mfi)* (p. 1-5). doi: 10.1109/MFI55806.2022.9913855

Liu, D., Smaragdis, P., & Kim, M. (2014). Experiments on deep learning for speech denoising. In *Interspeech.* Retrieved from https://api.semanticscholar.org/CorpusID:11587494

Loizou, P. C. (2012). *Speech enhancement: Theory and practice, second edition.* CRC Press. Retrieved from https://books.google.dz/books?id=ntXLfZkuGTwC&printsec=frontcover&redir_esc=y#v=onepage&q&f=false

Ludeña-Choez, J., & Gallardo-Antolín, A. (2012). Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints. In *Advances in speech and language technologies for iberian languages* (Vol. 328, pp. 207–216). Berlin, Heidelberg: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-35292-8_22 doi: 10.1007/978-3-642-35292-8_22

McLoughlin, I. (2016). *Speech and audio processing: A matlab-based approach.* Cambridge University Press. Retrieved from https://books.google.dz/books?id=r59ODAAAQBAJ&printsec=frontcover&redir_esc=y#v=onepage&q&f=false

Mel-cepstrum. (2021). Retrieved from https://speechprocessingbook.aalto.fi/Representations/Melcepstrum.html

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., … Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179-1210. doi: 10.1109/JSTSP.2022.3207050

Mohammadiha, N., Smaragdis, P., & Leijon, A. (2013). Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(10), 2140-2151. doi: 10.1109/TASL.2013.2270369

Murthi, M., & Rao, B. (1997). Minimum variance distortionless response (mvdr) modeling of voiced speech. In *1997 ieee international conference on acoustics, speech, and signal processing* (Vol. 3, p. 1687-1690 vol.3). doi: 10.1109/ICASSP.1997.598838

Nguyen, B., Uhlich, S., & Cardinaux, F. (2023). Improving self-supervised learning for audio representations by feature diversity and decorrelation. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1-5). doi: 10.1109/ICASSP49357.2023.10097100

of Music, D. M. (2024). Sound waves. *Johns Hopkins University*. Retrieved from `https://www.ams.jhu.edu/dan-mathofmusic/sound-waves/` (Accessed: 2024-11-12)

OpenStax. (2024). *13.2 wave properties: Speed, amplitude, frequency, and period.* Author. Retrieved from `https://openstax.org/books/physics/pages/13-2-wave-properties-speed-amplitude-frequency-and-period` (Accessed: 2024-11-12)

Popović, B., Krstanović, L., Janev, M., Suzić, S., Nosek, T., & Galić, J. (2022). Speech enhancement using augmented ssl cyclegan. In *2022 30th european signal processing conference (eusipco)* (p. 1155-1159). doi: 10.23919/EUSIPCO55093.2022.9909754

Processing, S. (2022). *Formula to calculate cepstral coefficients (not mfcc).* Retrieved from `https://dsp.stackexchange.com/questions/48886/formula-to-calculate-cepstral-coefficients-not-mfcc` (Accessed: 2024-11-12)

Rabiner, L. R., & Schafer, R. W. (2007). *Introduction to digital speech processing* (Vol. 1) (No. 1-2). Hanover, MA: now Publishers Inc. Retrieved from `https://research.iaun.ac.ir/pd/mahmoodian/pdfs/UploadFile_2643.pdf`

Removal, N. (2022). Noise removal in speech processing using spectral subtraction. , 20. Retrieved from `https://www.scirp.org/journal/paperinformation?paperid=45989`

Rocchesso, D. (2003). *Introduction to sound processing.* Università di Verona. Retrieved from `https://ia600309.us.archive.org/13/items/IntroductionToSoundProcessing/vsp.pdf`

Salamon, J., Jacoby, C., & Bello, J. P. (2014, Nov.). A dataset and taxonomy for urban sound research. In *22nd ACM international conference on multimedia (acm-mm'14)* (pp. 1041–1044). Orlando, FL, USA. Retrieved from `https://zenodo.org/records/1203745`

Sanada, Y., Nakagawa, T., Wada, Y., Takanashi, K., Zhang, Y., Tokuyama, K., … Yamada, T. (2022). Deep Self-Supervised Learning of Speech Denoising from Noisy Speeches. In *Proc. interspeech 2022* (pp. 1178–1182). doi: 10.21437/Interspeech.2022-306

Sivaram, T. (2024, March). Audio denoiser: A speech enhancement deep learning model. *Analytics Vidhya*. Retrieved from `https://www.analyticsvidhya.com/blog/2022/03/audio-denoiser-a-speech-enhancement-deep-learning-model/`

Spectral content (colour) of noise exposure affects work efficiency. (2024). *PMC*. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMCXXXXXX` (Accessed: 2024-11-29)

Spiceworks. (2020). *What is ml (machine learning)?* Retrieved from https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/amp

Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2020). Speech enhancement by lstm-based noise suppression followed by cnn-based speech restoration. *EURASIP Journal on Advances in Signal Processing*, *2020*(1), 49. doi: 10.1186/s13634-020-00707-1

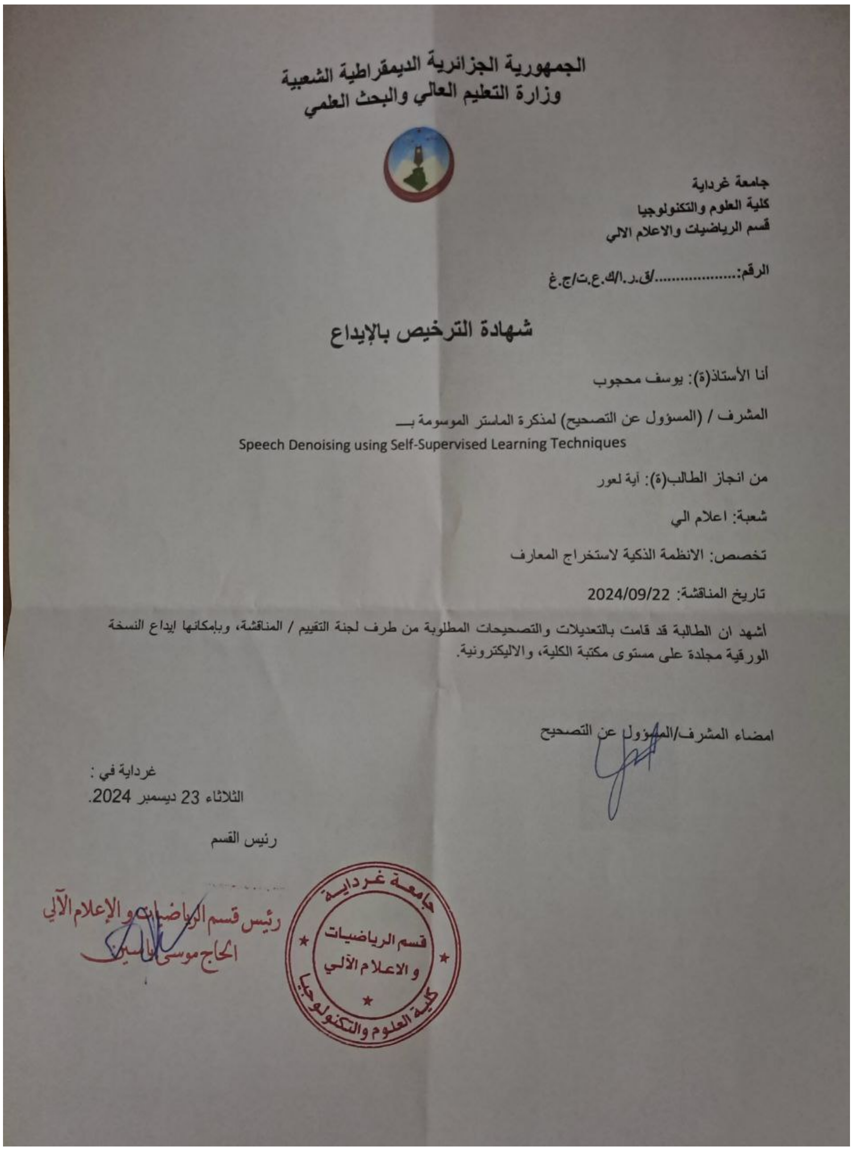
TechTarget. (2016). *Unsupervised learning.* Retrieved from http://www.techtarget.com/searchenterpriseai/definition/unsupervised-learning (Accessed on February 1, 2025)

Tektronix. (2024). *Spectrum view: A new approach to frequency domain analysis on oscilloscopes.* Retrieved from https://www.tek.com/en/documents/application-note/spectrum-view-new-approach-frequency-domain-analysis-oscilloscopes (Accessed: 2024-11-12)

Wilson, K. W., Raj, B., Smaragdis, P., & Divakaran, A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *2008 ieee international conference on acoustics, speech and signal processing* (p. 4029-4032). doi: 10.1109/ICASSP.2008.4518538

Wu, J., Li, Q., Yang, G., Li, L., Senhadji, L., & Shu, H. (2023). Self-supervised speech denoising using only noisy audio signals. *Speech Communication*, *149*, 63-73. Retrieved from https://hal.science/hal-04064230 doi: 10.1016/j.specom.2023.03.009

Feuille de correction



الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة غرداية
كلية العلوم والتكنولوجيا
قسم الرياضيات والاعلام الالي

الرقم:................../ق.ر.ا/ك.ع.ت/ج.غ

شهادة الترخيص بالإيداع

أنا الأستاذ(ة): يوسف محجوب

المشرف / (المسؤول عن التصحيح) لمذكرة الماستر الموسومة بـــ

Speech Denoising using Self-Supervised Learning Techniques

من انجاز الطالب(ة): آية لعور

شعبة: اعلام الي

تخصص: الانظمة الذكية لاستخراج المعارف

تاريخ المناقشة: 2024/09/22

أشهد ان الطالبة قد قامت بالتعديلات والتصحيحات المطلوبة من طرف لجنة التقييم / المناقشة، وبامكانها ايداع النسخة الورقية مجلدة على مستوى مكتبة الكلية، والاليكترونية.

امضاء المشرف/المسؤول عن التصحيح

غرداية في :
الثلاثاء 23 ديسمبر 2024.

رئيس القسم

رئيس قسم الرياضيات و الإعلام الآلي
الحاج موسى ياسين

جامعة غرداية
قسم الرياضيات
و الاعلام الآلي
كلية العلوم والتكنولوجيا

Figure 3.6: Enter Caption