Master's Thesis

Study on noise suppression based on the spectro-temporal modulation

Fanda Yuliana Putri

Supervisor Masashi Unoki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2023

Abstract

Noise suppression is an open-area research that addresses the challenge of obtaining speech from noisy speech as the result of environmental noise. The primary challenge of noise suppression is effectively attenuating the noise without substantially damaging the intricate balance between the quality and intelligibility of the noise-suppressed stimuli. Hence, achieving this delicate balance requires advanced consideration of the characteristic of speech and various noises in the respective features for noise suppression. An aggressive noise suppression may lead to the removal of important speech cues, affecting speech clarity and intelligibility in scenarios where speech and noise overlap in time; distinguishing between the two and suppressing noise while preserving speech components becomes more challenging.

In noise suppression algorithms that utilize spectral features, musical noise is a common problem. It refers to an undesirable artifact that occurs when certain noise components are mistakenly identified as speech and are suppressed inappropriately, resulting in a musical or tonal quality in the enhanced speech signal. This artifact often manifests as a periodic or rhythmic sound, resembling a musical note or a buzzing sound, which can be highly distracting and adversely affect speech intelligibility.

This study investigates a novel approach by integrating spectro-temporal modulation (STM) with a statistical noise suppression method, the minimum mean square error short-time spectral amplitude (MMSE-STSA) noise suppression algorithm, to achieve improvement in the noise suppression result. The research begins with a detailed exploration of the analysis-synthesis pipeline for STM feature extraction, highlighting the distinct spectro-temporal characteristics exhibited by speech and noise in the STM domain. Notably, the study delves into the averaged STM features of white, pink, and factory noises, revealing nuanced differences in their spectro-temporal properties, thus deepening an understanding of their impact on noise suppression.

The proposed analysis-modification-synthesis (AMS) pipeline is introduced, where the conventional noise suppression block is replaced by the implementation of the MMSE-STSA algorithm using the STM feature. This strategic integration leverages the joint spectro-temporal information provided by STM to adaptively obtain the estimation of speech, thereby enhancing the efficacy of the noise suppression algorithm. Moreover, based on empirical findings, the study uncovers the potential benefits of incorporating

over-suppression in the STM domain, leading to further improvement in noise reduction results. As a consequence, the MMSE-STSA algorithm undergoes modification to accommodate the over-suppression, enhancing its noise reduction capabilities.

Parameter tuning is conducted to optimize the attenuation gain (β) and enhance noise reduction, particularly in the speech-dominant groups of the STM domain. The study reveals specific β values that lead to better noise suppression results.

An evaluation of the efficacy of the proposed noise suppression algorithm utilizing the STM feature was conducted in comparison to established statistical noise suppression methods based on spectral features, namely the Wiener filter and the MMSE-STSA noise suppression algorithm. Three objective evaluation metrics, Segmental Signal-to-Noise Ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI), are utilized to assess the effectiveness of the proposed method.

The results indicate that the proposed algorithm achieves significant improvements in speech intelligibility, as demonstrated by higher STOI scores compared to other algorithms. However, the overall audio quality, as measured by PESQ, does not consistently surpass the benchmark methods. Despite this, the research identifies the potential of STM as an alternative feature for noise suppression, offering unique insights into the characterization of clean speech and various noises in the modulation domain.

Conclusively, the proposed STM-based noise suppression algorithm shows promise in enhancing speech intelligibility. While further research is needed to address certain limitations, such as the focus on single-channel uncorrelated noise and the non-linear nature of STM, the exploration of STM in noise suppression provides valuable contributions to this area of research. This study encourages future investigations to advance noise reduction algorithms and explore the potential of STM in various audio processing applications.

Keywords: noise suppression, temporal modulation, spectral modulation, spectro-temporal modulation, wiener filter, MMSE-STSA

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Research background

The noise suppression algorithm, as defined in the literature [3, 4], refers to techniques for eliminating disturbance or unwanted sound from an audio signal. The primary goal is to enhance both the speech quality as well as the speech intelligibility of the target audio by eliminating or mitigating background noise arising from diverse sources, including ambient sounds and electronic interference. In light of the escalating reliance on digital media in human activities, the demand for a robust noise suppression algorithm becomes increasingly evident. Its widespread application spans numerous domains, including audio recording, voice communication, speech recognition, and hearing aids, effectively addressing the need for pleasant auditory experiences in various real-world scenarios.

The degradation of the audio signal by noise has a detrimental effect on its quality, leading to challenges in comprehending the conveyed information [5]. Prolonged exposure to audio signals with diminished quality can induce unpleasant experiences, such as ear fatigue. Suppressing unwanted noise emerges as a viable solution, as it enhances the audibility of desired audio, facilitating more effective communication and elevating the overall listening experience. Particularly, in applications like voice communication, speech recognition, and hearing aids, noise interference can significantly impede the intelligibility of spoken words [6]. In these contexts, the noise suppression algorithm is commonly implemented as a front-end module preceding subsequent processing stages. The objective is to augment speech clarity and intelligibility by reducing background noise, facilitating more accessible communication, and accurate transcription of spoken content.

The existing statistical noise suppression methods are not without their challenges, as they may exhibit issues such as the occurrence of musical noise and various other types of suppression artifacts [3, 7]. These problems can be attributed to the limitations of commonly employed speech representations, such as waveform or spectral components. These representations do not

inherently facilitate an optimal separation of speech and noise, leading to inefficiencies in noise suppression. Consequently, novel approaches that address these limitations and offer improved speech-noise discrimination are sought to enhance the efficacy and reliability of noise suppression techniques.

1.2 Research issues

The research domain of noise suppression encompasses several significant challenges. A primary concern pertains to noise estimation and modeling, which holds critical importance in the development of effective noise suppression algorithms [8,9]. Accurately estimating noise in real-world environments is challenging due to varying acoustic conditions and the overlap of speech and noise signals. This overlap presents difficulties in effectively separating desired speech from background noise while preserving speech characteristics. Addressing this issue involves developing advanced algorithms to better distinguish and model speech and noise components for more effective noise suppression.

The next issue is maintaining the amount of noise reduction without sacrificing speech quality [3, 10]. This is essential to develop a noise suppression algorithm to attain a balance between reducing noise and preserving speech characteristics. Excessive noise reduction can result in speech distortion, leading to reduced speech intelligibility or compromised audio quality. The relative significance of speech quality and intelligibility depends on the use case. For instance, optimizing speech quality is preferred for enhancing the listening experience of human listeners. Conversely, applications like speech recognition or hearing aids prioritize speech intelligibility to ensure effective communication and accurate processing [11, 12]. Consequently, the development of algorithms capable of reducing noise effectively while mitigating speech distortion poses an ongoing challenge within this research domain.

1.3 Research motivation

Statistical noise suppression methods primarily operate in the spectral domain, encompassing techniques like spectral subtraction [13], Wiener filtering [10, 14, 15], and minimum mean-square error (MMSE) [8]. More recently, the ideal ratio mask (IRM) using deep learning models has been explored and achieved a satisfactory result [16]. Despite their effectiveness, it has been noted that uncontrolled noise suppression algorithms may yield noise artifacts

worse than the actual noise [17]. Hence, addressing this issue becomes pivotal in advancing noise suppression methodologies.

The modulation feature, extensively employed in noise suppression, has been widely investigated in speech research [4]. Early investigations in the modulation domain characterized speech signals as audible carriers and amplitude modulation signals (AM) [18]. Temporal modulation pertains to changes in the temporal envelope of stimuli. Notably, speech exhibits lower modulation components with greater depth compared to most noise-like signals. Research has revealed that speech information primarily resides in slower temporal modulations, predominantly around 3-4 Hz [19]. Leveraging this acoustic foundation, noise reduction algorithms have been developed to distinguish speech from noise using the temporal modulation feature. In the context of noise suppression studies, the implementation of temporal modulation features using diverse noise suppression techniques has demonstrated promising outcomes, including spectral subtraction [20] and MMSE [21]. Both studies reported diminished instances of musical noise compared to their spectral domain counterparts. Additionally, spectral modulation provides insights into the periodicity of spectral components, similar to cepstrum analysis frequently employed for source-filter separation [22–24].

The spectro-temporal modulation (STM) pertains to changes in the joint spectro-temporal envelopes of the signal. Extensive research has employed the joint STM feature to investigate the mammalian auditory system [25,26]. The significance of temporal, spectral, and joint spectral-temporal modulation in speech perception has been thoroughly examined [2,27,28]. Moreover, a separate study demonstrates the feasibility of distinguishing noise and speech using this feature [29]. Despite these findings, the application of spectro-temporal modulation in noise suppression remains relatively limited.

1.4 Research objectives

The primary objective of this research is the development of a noise suppression algorithm characterized by enhanced resilience to suppression artifacts. One approach to achieving this goal involves the exploration of an alternative feature, which can yield improved differentiation between speech and noise. The feature under investigation in this research is spectro-temporal modulation, which has been empirically demonstrated to offer superior discrimination between speech and noise.

The novelty of the proposed method is in its utilization of the spectro-temporal modulation domain, an area that has seen limited application in noise suppression despite the promising separability between noise and

speech. A comprehensive exploration of various noise and speech characteristics in the modulation domain provides valuable insights for future noise suppression research. Furthermore, this research evaluates an extension of the statistical signal processing algorithm using the spectro-temporal modulation feature, offering a novel approach to noise suppression that capitalizes on the distinctive properties of the modulation domain.

The coverage of this study is constrained to the exploration and implementation of a noise suppression algorithm focused on single-channel noise suppression scenarios involving uncorrelated noise and speech information, such as additive noise.

1.5 Organization of thesis

The thesis is structured into six chapters, with each chapter comprising the following details.

- Chapter 1 introduces the research background and motivation, focusing on noise suppression through spectro-temporal modulation. The section addresses key issues within the research area, outlines the objectives and novelty of the study, and provides the organization of the thesis.
- Chapter 2 is the literature review containing the basic and necessary information for noise suppression, including several classical techniques on noise suppression in the spectral domain and the modulation domain. This chapter also covers the description of the different approaches to extracting the spectro-temporal modulation features.
- Chapter 3 covers the detailed definition of the proposed method. The subsection includes the synthesis and analysis of the modulation feature, the defining characteristics of various types of speech and various noises in the modulation domain, and the proposed noise suppression algorithm.
- Chapter 4 contains the description of the dataset and the parameter tuning. In the parameter tuning subsection, the estimation of the attenuation gain as well as the modulation frequencies partition, is explained.
- Chapter 5 covers the result of the evaluations, including the description of the evaluation metrics, the noise suppression result, and the general discussion.
- Chapter 6 concludes the thesis by describing the research summary, remaining works, and contributions.

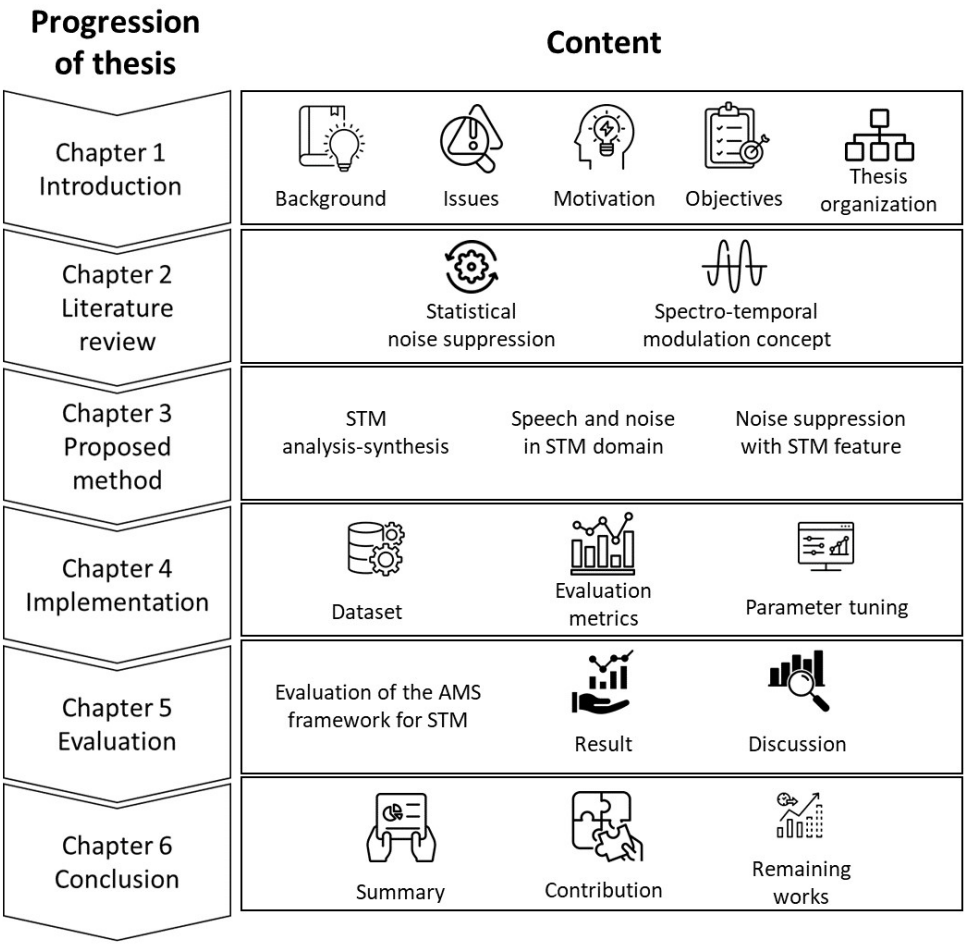


Figure 1.1: Organization of the thesis

Chapter 2

Literature review

2.1 Overview of the statistical noise suppression

This section provides an overview of statistical noise suppression algorithms, encompassing spectral subtraction, Wiener filtering, and the MMSE-STSA. These algorithms, elucidated within this section, operate using the frequency or spectral features, achieved from performing the Fourier transform (FT) on the signal.

2.1.1 AMS framework using spectral feature

In general, the algorithms explained in this chapter adhere to the analysis-modification-synthesis (AMS) pipelines, depicted in Fig. 2.1, where the analysis step involves applying the short-time Fourier transform (STFT) to the noisy signal. The synthesis step entails utilizing the inverse STFT to reconstruct the estimated clean signal. The modification step may incorporate one of the noise suppression algorithms. Consequently, this subsection focuses on elaborating on the analysis and synthesis steps. For more comprehensive information on the modification process, further details are available at [30].

2.1.1.1 Analysis

The Analysis step entails performing the STFT, including windowing the input signal by segmenting it to form overlapping frames and applying the FT. To elaborate on the analysis step in its entirety, the process is described in the following equation.

$$X(l,\omega_k) \triangleq X(l,k) = \sum x(n)w(l-n)e^{-j\frac{2\pi}{M}kn}\,, \tag{2.1}$$

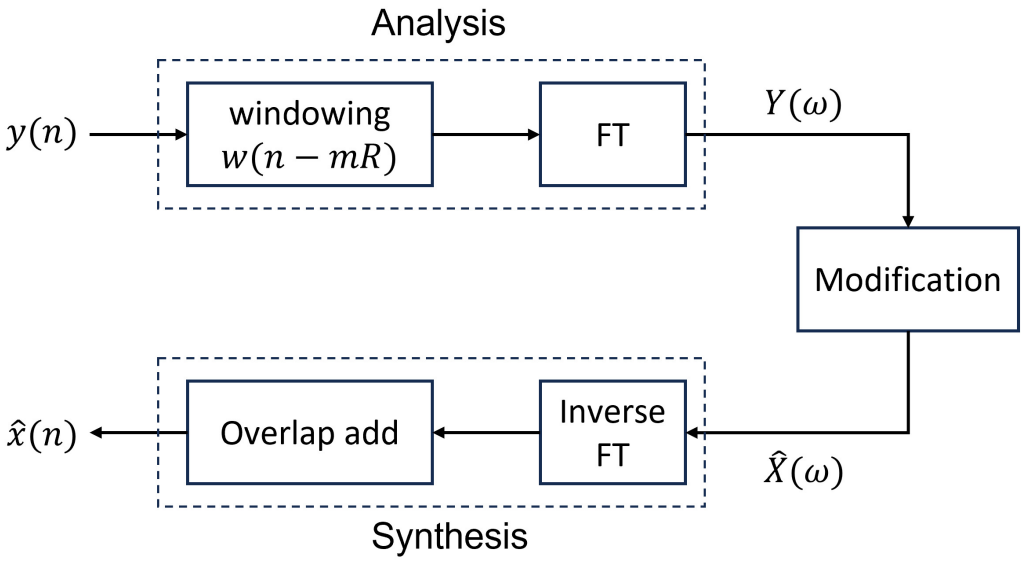


Figure 2.1: AMS framework using the spectral feature.

with $x(n)$ as the input and the $w(n)$ as the analysis window. The l is the index for the discrete time, and k is the index for the frequency. M is the samples in a window that defines the spacing in the frequency axis, such that the w_k is the M uniformly sampled frequencies, i.e., at $w_k = 2\pi k/M, k = 0, 1, ..., M-1$.

Several windowing functions are commonly used for signal processing and spectral analysis. One of them is the Hamming window [31]. This window function tapers the edges of the signal segment to minimize spectral leakage.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) , \tag{2.2}$$

where $w(n)$ represents the value of the Hamming window at index n which is the index of the window ranging from 0 to $N-1$. N represents samples of the window.

2.1.1.2 Synthesis

One of the methods to generate the reconstructed signal from its STFT is known as the inverse STFT or just ISTFT defined as follows:

$$x(n) = \frac{1}{M} \sum S(l,k) w^*(l-n) e^{j\frac{2\pi}{M}kn} , \tag{2.3}$$

where $x(n)$ is the reconstructed discrete signal and $w^*(l-n)$ is the complex conjugated window function.

2.1.2 Spectral subtraction

This noise suppression algorithm is among the earliest proposed noise suppression algorithms and has been subject to various studies aiming to enhance its performance. Initially implemented using spectral features, this algorithm operates on a straightforward basis of computation. By assuming the case of noisy speech augmented with additive noise, the subtraction of the estimation of the noise spectrum from the noisy spectrum is computed to achieve the estimation of the clean spectrum. During short intervals of speech absence, the calculation and update of the noise spectrum are done assuming the noise signal varies slowly (i.e., exhibits a more prominent stationary tendency) in comparison to the speech, which is present closely in an intermittent manner. Subsequently, the estimated or enhanced audio is generated by the inverse FT of the estimated clean spectrum, incorporating the noisy phase information.

Spectral subtraction is renowned for its susceptibility to a distortion issue termed "musical noise." This peculiar type of noise typically manifests as a residual artifact of the noise suppression process, characterized by a tonal or musical-like quality that deviates from the original signal or the intended noise reduction. To address this problem, several research endeavors have proposed strategies to alleviate or, in certain instances, entirely eliminate the presence of musical noise in this noise suppression algorithm.

2.1.2.1 Basic principle

The algorithm assumes $y(n)$, the noisy input degraded by additive noise, consists of the speech $x(n)$ and the noise signal $d(n)$ with additive relation, as demonstrated below.

$$y(n) = x(n) + d(n)\,. \tag{2.4}$$

Applying the discrete Fourier transform (DFT) of both sides leads to:

$$Y(\omega) = X(\omega) + D(\omega)\,. \tag{2.5}$$

The result of the DFT, such as $Y(\omega)$, is a complex number that can be expressed in the polar form.

$$X(\omega) = |X(\omega)|e^{j\phi_x(\omega)}\,, \tag{2.6}$$

where $|X(\omega)|$ is the spectral amplitude and $\phi_x(\omega)$ is the spectral phase of the speech. The noise and noisy signals are expressed similarly.

Similar to the general scenario of single-channel noise suppression, the sole available information is the noisy signal $y(n)$. In spectral subtraction,

the clean spectral amplitude estimation is accomplished by the subtraction of the noisy spectral amplitude and the estimated noise spectral amplitude, which is represented as follows:

$$|\hat{X}(\omega)| = \max\{|Y(\omega)| - E\left[|D(\omega)|\right], 0\}, \tag{2.7}$$

where $|\hat{X}(\omega)|$ is the estimated clean spectral amplitude. Meanwhile, the phase information of the estimated clean spectrum is obtained from the noisy spectrum, i.e., $\angle\hat{X}(\omega) = \angle Y(\omega)$. This is driven by the observation that phase information has a minimum impact on speech intelligibility [32].

2.1.2.2 Power spectral subtraction

The generalized version of the Eq. (2.7) is defined as follows:

$$|\hat{X}(\omega)|^{\alpha} = |Y(\omega)|^{\alpha} - E\left[|D(\omega)|^{\alpha}\right]. \tag{2.8}$$

Hence, the power spectral subtraction can be defined with $\alpha = 2$.

$$|\hat{X}(\omega)|^{2} = |Y(\omega)|^{2} - E\left[|D(\omega)|^{2}\right]. \tag{2.9}$$

Another way to express the equation is by defining the gain of the spectral subtraction as follows:

$$H(\omega) = \frac{|\hat{X}(\omega)|}{|Y(\omega)|} = \sqrt{\frac{|Y(\omega)|^2 - E[|D(\omega)|^2]}{|Y(\omega)|^2}}. \tag{2.10}$$

2.1.2.3 Over-subtraction

To mitigate the prevalent issue of musical noise in spectral subtraction, the over-subtraction technique has been introduced [33]. This problem is primarily caused by inaccurate noise information estimation and the application of half rectification, as depicted in Eq. (2.7).

The logic behind the over-subtraction technique is to reduce the noise peaks by filling the gaps or valleys at specific frequencies to mask the residue noise, as illustrated by Fig. 2.2.

$$|\hat{X}(\omega)|^{\alpha} = \max\{|Y(\omega)|^{\alpha} - \beta E\left[|D(\omega)|^{\alpha}\right], \delta E\left[|D(\omega)|^{\alpha}\right]\}, \tag{2.11}$$

where β and δ are the attenuation gain that can be controlled. However, it is common to apply $\beta > 0$ and $0 < \delta \ll 1$.

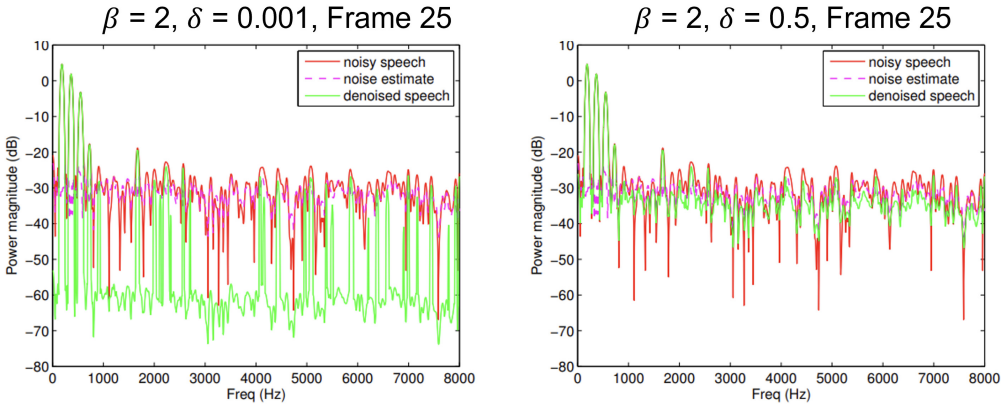


Figure 2.2: Over-subtraction in spectral subtraction. Figure retrieved from [1]

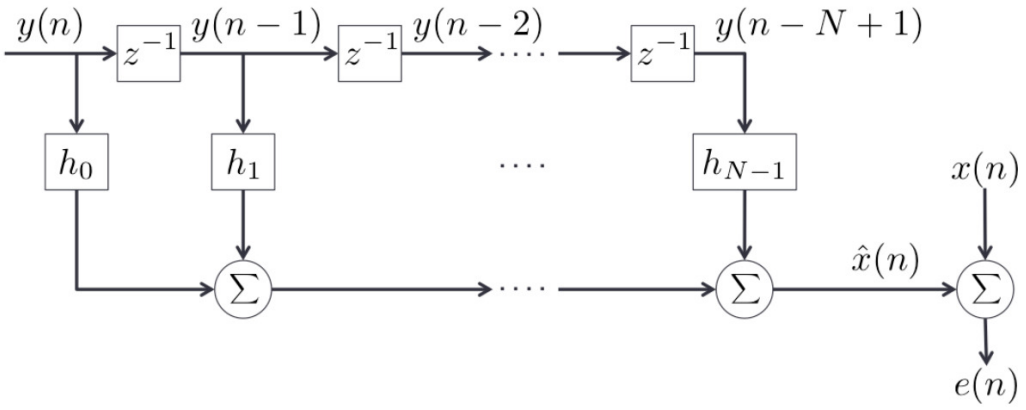


Figure 2.3: Wiener filter. Figure retrieved from [1]

2.1.3 Wiener filter

The Wiener filter is a prominent filter commonly applied in signal processing and image restoration tasks, it estimates the clean speech by calculating the MMSE between the estimated signal and the reference signal. This filter utilizes information regarding speech and noise statistics in the frequency domain to design an optimal filter. Similar to the previously defined spectral subtraction algorithm, the operation of the Wiener filter is done assuming the additivity and stationary tendency from the noise and speech.

In [14, 15], the Wiener filter is defined as finding the optimal linear filter that outputs the estimated clean signal as described by

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{a=0}^{M-1} h(n)y(n-a)\,, \tag{2.12}$$

with $e(n)$ is the residue and $h(n)$ is the linear filter. In case of a non-causal infinite impulse response (IIR) filter, using the convolution theorem, i.e., $x(n) * h(n) \leftrightarrow X(\omega)H(\omega)$, then Eq. (2.12) is defined as follows:

$$E(\omega) = X(\omega) - H(\omega)Y(\omega)\,. \tag{2.13}$$

Then, by minimizing $E[|X(\omega)|^2]$ limited by $H(\omega)$, the general equation of the Wiener filter is defined as follows:

$$H(\omega) = \frac{E[|X(\omega)|^2]}{E[|Y(\omega)|^2]} = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E[|D(\omega)|^2]}\,, \tag{2.14}$$

with $H(\omega)$ is also known as the gain of the Wiener filter.

In single-channel noise suppression, $E[|X|^2]$ is unknown. However, it can be estimated as follows:

$$E[|X(\omega)|^2] = |Y(\omega)|^2 - E[|D(\omega)|^2]\,. \tag{2.15}$$

Substitution of Eq. (2.15) to Eq. (2.14), the gain is obtained as follows:

$$H(\omega) = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E[|D(\omega)|^2]} = \frac{|Y(\omega)|^2 - E[|D(\omega)|^2]}{|Y(\omega)|^2}\,. \tag{2.16}$$

The estimated clean spectral amplitude is calculated by multiplying the gain with the spectral amplitude of the noisy signal. Notice that from Eq. (2.10) and Eq. (2.16), the square root Wiener filter is equal to the power spectral subtraction.

The reconstruction of the estimated clean signal is generated by substituting the noisy spectral phase for the estimated clean signal. In this case, the Wiener filter also assumes a similar assumption as the spectral subtraction, in which the phase information does not significantly affect speech intelligibility.

2.1.4 MMSE-STSA

MMSE-STSA algorithm is established and well-known to be a robust noise suppression algorithm. This algorithm was proposed in [8] and the log-MMSE, which was proposed in [34].

The MMSE-STSA algorithm has several key assumptions. These assumptions are idealized models and may not hold in all real-world scenarios. The

performance of the MMSE-STSA algorithm can varies on the extent to which these assumptions are true.

- Additive Gaussian noise. The additive characteristic of the noise to the speech signal is assumed to be stationary and follows the Gaussian distribution.
- Time-invariant speech absence probability (SAP). It is assumed that the SAP in each frequency bin remains constant throughout the short-time segments. This assumption calculates the estimation of the SAP accurately and hence is generally utilized to determine the gain function for speech enhancement.
- Uncorrelated speech and noise. This algorithm assumes uncorrelatedness and independence between speech and various noise signals.

The basic idea behind MMSE-STSA is to calculate the estimation of the clean speech from the noisy stimuli by calculating the MMSE from the noise-suppressed and the reference signal. Hence, given $X(\omega) = Xe^{j\angle X(\omega)}$,

$$\min\{E[(X(\omega) - \hat{X}(\omega))^2]\}\,. \tag{2.17}$$

From the Bayesian statistics, the optimal MMSE estimator is defined as

$$\begin{aligned}\hat{X}(\omega) &= E[X(\omega)|Y(\omega)] \\ &= \int_0^\infty X(\omega)P(X(\omega)|Y(\omega))dX \\ &= \frac{\int_0^\infty X(\omega)p(Y(\omega)|X(\omega))p(X(\omega))dX}{p(Y(\omega))}\,.\end{aligned} \tag{2.18}$$

From Eq. (2.18), we need to know the distribution of $X(\omega)$ and $X(\omega)$, which is assumed to be Gaussian as mentioned in the list of assumptions above. More specifically, this algorithm assumed that the FT coefficient (of both noise and speech) is Gaussian. As a result, based on the central limit theorem (CLT), which states that independent and identically distributed (i.i.d.) random variables are added together, their sum tends to follow a normal (Gaussian) distribution, the noisy signal is also assumed to follow the same distribution. This is important to be noted as the CLT also holds for weakly dependent signals. However, the variance of the distribution $E[Y(\omega)]$ is time-varying. Below is the summary

$$\begin{aligned}X(\omega) &\sim \mathcal{N}(0, E[|X(\omega)|^2]) \\ D(\omega) &\sim \mathcal{N}(0, E[|D(\omega)|^2]) \\ Y(\omega) &\sim \mathcal{N}(0, E[|X(\omega)|^2] + E[|D(\omega)|^2])\,.\end{aligned} \tag{2.19}$$

The definition of spectral gain can be represented with two variables, the A Priori SNR as well the A Posteriori SNR, which are described below, respectively.

$$\xi(\omega) = \frac{E[|X(\omega)|^2]}{E[|D(\omega)|^2]}\,, \tag{2.20}$$

$$\gamma(\omega) = \frac{|Y(\omega)|^2}{E[|D(\omega)|^2]}\,. \tag{2.21}$$

Using some temporary variable, $\nu(\omega)$

$$\nu(\omega) = \frac{\xi(\omega)}{1+\xi(\omega)}\gamma(\omega)\,. \tag{2.22}$$

The gain function of the MMSE-STSA algorithm is described in the equation below:

$$H(\xi,\gamma) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu}}{\gamma}\exp\left(-\frac{\nu}{2}\right)\left[(1+\nu)I_0\left(\frac{\nu}{2}\right)+\nu I_1\left(\frac{\nu}{2}\right)\right]. \tag{2.23}$$

Hence, the estimated clean spectral amplitude can be obtained by multiplying the MMSE-STSA gain with the noisy spectral amplitude defined below.

$$|\hat{X}(\omega)| = H(\xi,\gamma)\cdot|Y(\omega)|\,. \tag{2.24}$$

2.1.4.1 Decision directed approach

Notice that in Eq. (2.20), the clean spectral amplitude is not available in the single channel noise suppression method. The A Priori SNR is estimated similarly to the derivation of the Wiener filter by Eq. (2.15). The newly estimated A Priori SNR is:

$$\hat{\xi}(\omega) = \frac{|Y(\omega)|^2 - E[|D(\omega)|^2]}{E[|D(\omega)|^2]} = \frac{|Y(\omega)|^2}{E[|D(\omega)|^2]} - 1\,. \tag{2.25}$$

The decision-directed approach is proposed in the study to obtain the estimation of the A Priori SNR:

$$\hat{\xi}(l,\omega) = \alpha\frac{|\hat{X}^2(l-1,\omega)|}{E[|D(l-1,\omega)|^2]} + (1-\alpha)\left(\frac{|Y(l,\omega)|^2}{E[|D(l,\omega)|^2]} - 1\right). \tag{2.26}$$

Similar to the two mentioned noise suppression algorithms, the estimated clean signal is obtained by applying ISTFT to the estimated clean spectral amplitude with a noisy spectral phase.

2.2 Spectro-temporal modulation concepts

2.2.1 Joint spectro-temporal modulation

Extensive research has been conducted on the spectro-temporal modulation feature to obtain an understanding of the underlying mechanisms of human perception and auditory processing [2, 25]. In a study by Elliott and Theunissen [2], the modulation transfer function (MTF) was investigated with a specific focus on speech intelligibility by using the modulation power spectrum (MPS). The MPS utilized in this study represents the joint STM.

The researchers conducted perceptual experiments involving human listeners to measure speech intelligibility under different conditions. The speech signals are manipulated by applying band-pass filtering at various modulation frequencies. During the analysis, the pattern observed in the modulation domain obtained from the original speech and the filtered versions are compared, in which the comparison result is used to assess the leverage induced by different modulation frequencies on speech intelligibility.

The study involved presenting the participants with modified audio signals to assess their ability to identify and comprehend the content. The investigation focused on the changes in the MTF, which is influenced by modulation frequencies. These changes were measured and compared to the speech comprehension skills of the participants.

The integrated experimental and computational approach in this study aimed to characterize the MTF in regard to speech intelligibility. By investigating the role of temporal modulations in speech perception, the research provided valuable insights into the critical impact of different modulation frequencies on speech intelligibility. The findings contribute significantly to the understanding of the way temporal modulations influence speech perception and offer a framework for studying and optimizing speech communication systems.

In conclusion, the study highlights the role of low modulation frequencies in both the temporal and spectral domains for speech intelligibility. Notably, the sense of perception was notably degraded as temporal modulations at 12 Hz or spectral modulations at 4 cyc/kHz were erased. The MTF demonstrated a band-pass characteristic in temporal modulations, ranging from 1 to 7 Hz, and a low-pass characteristic in spectral modulations, particularly at 1 cyc/kHz. These frequency ranges were identified as the most crucial for speech intelligibility.

For a detailed visual representation, please refer to Fig. 2.4, where the findings are illustrated.

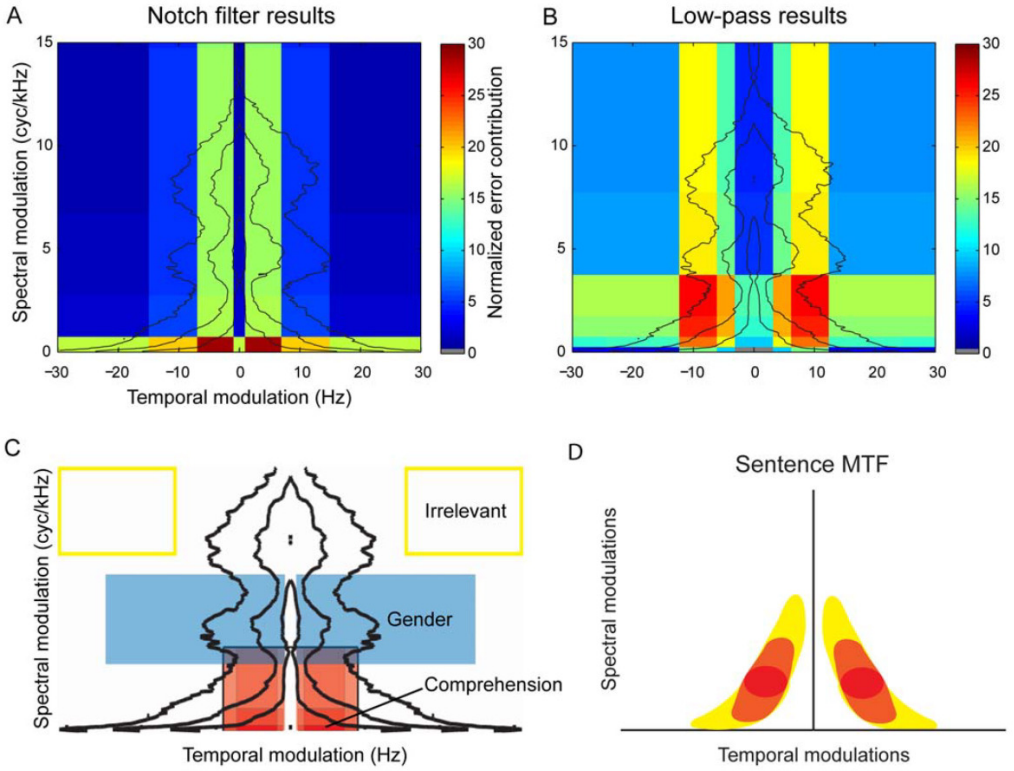


Figure 2.4: MTF-based on the spectro-temporal feature. Figure retrieved from [2]

2.2.2 Multi-resolution spectro-temporal modulation

Another group of researchers has explored the application of the multi-resolution spectro-temporal modulation, originally introduced in [25]. The paper presents a model for computation based on the multiple stages in the auditory analysis that yields a 4-dimensional multi-resolution representation of joint STM features in complex sounds. In simpler terms, the technique produces multiple spectrograms by filtering the spectral feature (in this case, the auditory spectrogram) of the signal based on various sets of temporal and spectral modulations.

This representation of spectro-temporal modulation has served as the fundamental feature in several other studies, including speech intelligibility prediction [35, 36], voice activity detection [37], speech enhancement [29], input feature for automatic speech recognition (ASR) [38], and as the loss-function for neural networks [39].

In the previously described studies, the multi-resolution STM feature is not a linear feature, as it utilizes an auditory-based spectral representation.

Consequently, a linear version of a similar feature is proposed in [40], and this feature has also been used as a feature for speech enhancement, as reported in [41].

The multi-resolution STM feature discussed in this section differs from the STM feature used in this study. While the multi-resolution STM employs a 4-dimensional feature comprising the filtered spectrogram based on modulation frequency, the STM feature utilized in this study focuses on the weights of the joint spectral-temporal modulations. Further details of the STM feature used in this study are elucidated in the subsequent chapters.

2.2.3 Noise suppression with modulation feature

Various studies have investigated noise suppression methods using modulation techniques, particularly focusing on temporal modulation. For instance, as described in [20], the modulation domain was explored as an alternative to the spectral domain for noise suppression using spectral subtraction. The modulation spectral subtraction algorithm was introduced and could effectively compensate for additive noise distortion, which as a result, outperformed the MMSE method. The experiment demonstrated consistent superiority of the proposed method in all SNR ranges, with notable improvements observed for lower SNR levels.

The study presented in [21] explored the enhancement of speech by studying MMSE short-time spectral magnitude estimation using the temporal modulation feature. The proposed approach effectively enhanced the quality of the processed signal without introducing musical noise or spectral smearing distortion. Notably, the mean subjective preference scores obtained for the proposed approach were significantly higher than those of other enhancement methods. This finding suggests that listeners perceived the proposed approach to exhibit superior quality compared to other methods under evaluation.

Chapter 3

Proposed method

3.1 Speech analysis-synthesis on STM-based feature

3.1.1 Analysis

The analysis step primarily is extracting the STM feature. The detailed step-by-step process for obtaining this feature is illustrated in Fig. 3.1.

The input of the analysis process is an audio signal. The first step is the spectrogram extraction, as in the analysis process explained in Section 2.1.1.1. Using simple explanation, the spectrogram extraction contains steps such as segmentation, windowing, Fourier transform, amplitude calculation, and an optional amplitude scaling (e.g., log-scaling). In this study, the window size is 20 ms with a window shift equal to 50% of the window size. Meanwhile, the number of FFT bins is 320. The spectrogram provides valuable insights into how the signal frequency components change with respect to time, revealing variations in spectral characteristics and highlighting time-varying patterns. The illustration of the spectrogram is in Fig. 3.1.

The second step involves short-time spectrogram extraction. Within this step, the previously obtained amplitude spectrogram undergoes further

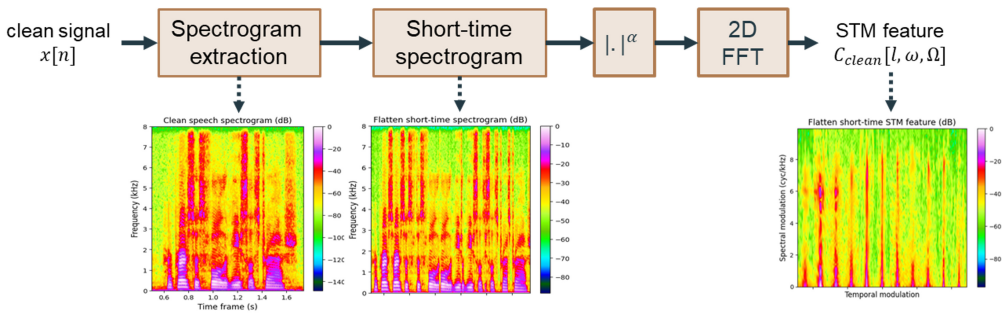


Figure 3.1: STM analysis

segmentation in overlapping blocks, employing a modulation block size of 100 ms and a modulation block shift of 50%. For a visual representation, the illustration of the short-time spectrogram is also available in Fig. 3.1.

The third step is the application of the power function to the amplitude spectrogram, $|.|^{\alpha}$. In this case, the value of α can be 1 or 2. Subsequently, if α equals 2, the output spectrogram is referred to as the short-time power spectrogram. In this study, α of 1 is used, which means the short-time amplitude spectrogram is utilized to obtain the STM feature.

The final step involves the application of the 2D windowing function and the 2D-FFT function to each short-time amplitude spectrogram block, yielding the conclusive STM feature. The specific windowing function employed in this study is the 2D Hamming window, as defined in the equation provided below. By meticulously processing the spectrogram in a block-by-block manner with the Hamming window and subsequently performing the 2D-FFT, the resultant STM feature encapsulates crucial information pertaining to the joint STM within the signal.

This approach ensures a comprehensive and coherent analysis of the joint spectro-temporal dynamics of the spectral feature. Hence, the feature for this study is called the short-time joint STM feature.

$$w(l,k) = 0.54 - 0.46\cos\left(\frac{2\pi l}{L-1}\right)\cos\left(\frac{2\pi k}{K-1}\right), \tag{3.1}$$

$w(l,k)$ is the 2D Hamming window at index l, and k. L and K are the width and the height of the modulation block, which are 10 and 5, respectively. The 2D-FFT is described as follows.

$$C_x(l,\omega,\Omega) = \sum\sum X(l,k)\cdot w(l-\Omega,k-\omega)\cdot e^{-j2\pi\left(\frac{\Omega l}{L}+\frac{\omega k}{K}\right)}. \tag{3.2}$$

The result of applying the 2D-FFT is a complex-valued matrix encompassing essential amplitude and phase information at distinct modulation frequencies, including temporal and spectral domains. The spectrogram, attained with the specified parameter configuration, yields an STM feature with temporal modulation spanning from -50 to 50 Hz. Additionally, the spectral modulation ranges from 0 to 10 cyc/kHz, characterizing the spectral variations across the analyzed signal. These ranges delineate the scope of the STM feature and its pivotal role in capturing the spectral and temporal dynamics inherent in the signal, thus contributing to a comprehensive understanding of the analyzed acoustic properties.

3.1.2 Synthesis

The synthesis process involves the application of the inverse counterparts of the processes explained in the Analysis step. Starting by the integration of the amplitude with the phase value derived from the STM feature, the subsequent stage entails the application of the inverse 2D-FFT. Subsequently, the overlap-add method is employed to obtain the reconstruction of both the spectrogram and the signal effectually.

By retracing the steps in reverse order, this process restores the original signal from its corresponding STM representation, resulting in the accurate reconstruction of the spectrogram and the ultimate audio signal.

3.2 Noise and speech in the STM-domain

Initially, the STM feature comprises a 3-dimensional matrix. Nonetheless, for the sake of facilitating a more accessible analysis, the presentation will be streamlined by focusing solely on the averaged STM feature. This pragmatic approach allows for a concise representation while retaining the essential information required to achieve the intended analytical objectives.

3.2.1 Speech

Fig. 3.2 displays the averaged STM feature of human speech, offering valuable insights into its spectro-temporal properties. The left figure represents the flattened STM feature obtained from the short-time spectrogram, while the right figure exhibits the averaged STM feature. Remarkably, the energy distribution in human speech concentrates significantly on frequencies in the lower temporal modulation ($\pm$ 20 Hz) and frequencies in the lower spectral modulation (below 2 cyc/kHz). This observation reinforces the consistency and congruence with the STM figure presented in a previous study [26], validating the significance and relevance of our current findings.

According to the analysis of the averaged spectral modulation depicted in Fig. 3.3, discernible peaks emerge below 2 cyc/kHz and around 6-8 cyc/kHz. These distinct peaks signify the source-filter information, distinguishing the fundamental frequency (F0) as well as the characteristics of the vocal tract, respectively. Notably, from the left figure in Fig. 3.3, male speakers exhibit a higher peak, indicative of lower F0 values, whereas female speakers manifest a lower peak, signifying higher F0 values. These demonstrate the salient vocal characteristics between female and male speakers.

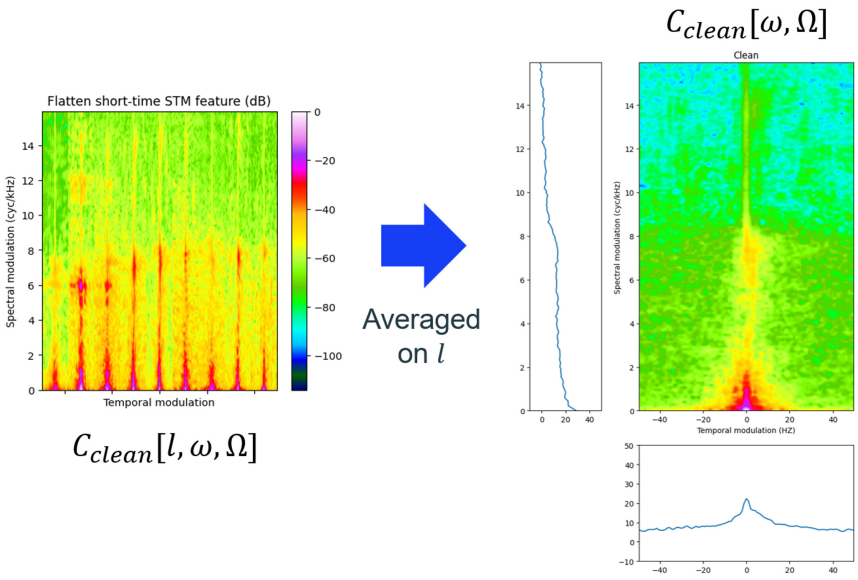


Figure 3.2: Averaged STM feature of human speech.

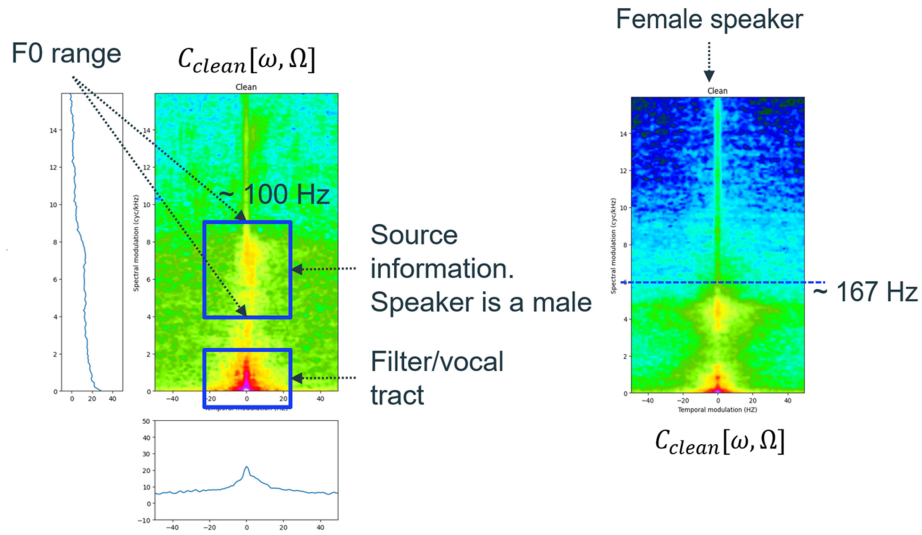


Figure 3.3: The source filter information of human speech in STM feature

3.2.2 Noise

3.2.2.1 White noise

Fig. 3.4 displays the averaged STM feature of white noise. The upper two figures, arranged from left to right, portray the spectrogram and the short-time STM feature, respectively. Meanwhile, the lower figure showcases the averaged STM feature, revealing distinct features unique to white noise.

Notably, the white noise exhibits a pronounced peak around zero modulation frequencies, both in the temporal and spectral domains, indicative of its centralized frequency distribution. Additionally, a prominent display of flat noise energy is observed in higher modulation frequencies. These findings define the intrinsic spectro-temporal properties of white noise and underscore its significance in acoustic analysis and noise suppression implementations.

3.2.2.2 Pink noise

Fig. 3.5 displays the averaged STM feature of pink noise. The upper two figures, arranged from left to right, portray the spectrogram and the short-time STM feature, respectively, providing a comprehensive overview of the acoustic attributes of the noise. Meanwhile, the lower figure showcases the averaged STM feature, revealing distinct features unique to pink noise.

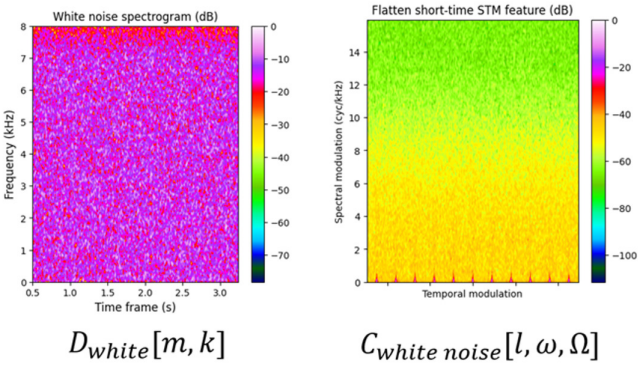
In the STM domain, the characteristics of pink noise closely resembles white noise, with the main distinction lying in the higher energy observed in the lower temporal modulation frequencies along all the spectral modulation. This heightened energy in the lower temporal modulation translates to a relatively higher concentration of energy in that range. Consequently, the energy distribution across other temporal modulation frequencies appears flatter in comparison to white noise, exhibiting a diminished ratio. This distinction indicates the nuanced differences in the spectro-temporal properties of pink noise, accentuating its relevance in acoustic analysis and its distinct impact on noise suppression techniques compared to white noise.

3.2.2.3 Factory noise

Fig. 3.6 displays the averaged STM feature of factory noise. The upper two figures, arranged from left to right, portray the spectrogram and the short-time STM feature, respectively, providing a comprehensive overview of the acoustic attributes of the noise. Meanwhile, the lower figure showcases the averaged STM feature, revealing distinct features unique to factory noise.

Within the STM domain, the attributes of factory noise bear a close resemblance to those of pink noise, yet a discernible discrepancy manifests

in the fluctuating energy observed within the higher temporal modulation frequencies. Notably, the ratio of flat energy for the higher temporal modulation exhibits a greater prominence in the context of factory noise.



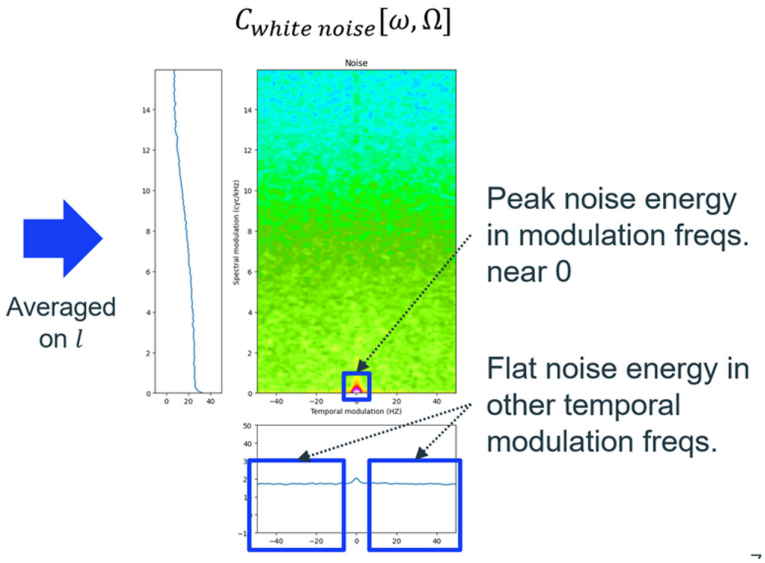


Figure 3.4: Sample of STM feature of white noise

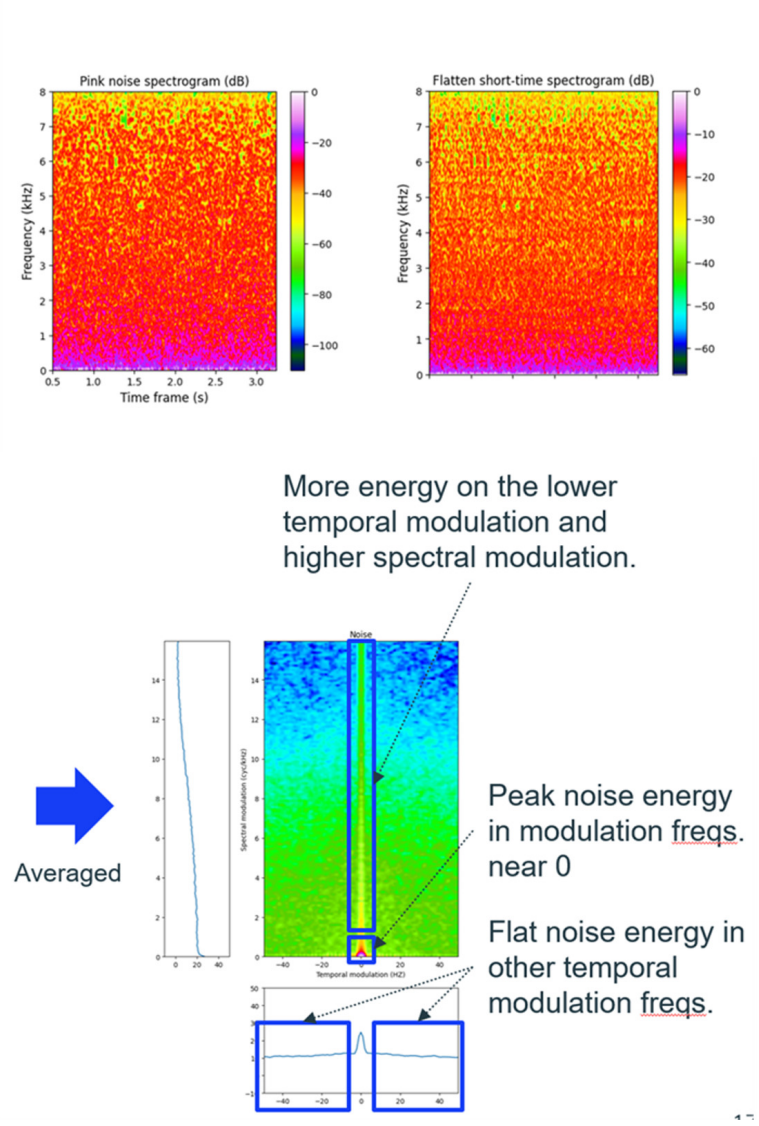


Figure 3.5: Sample of STM feature of pink noise

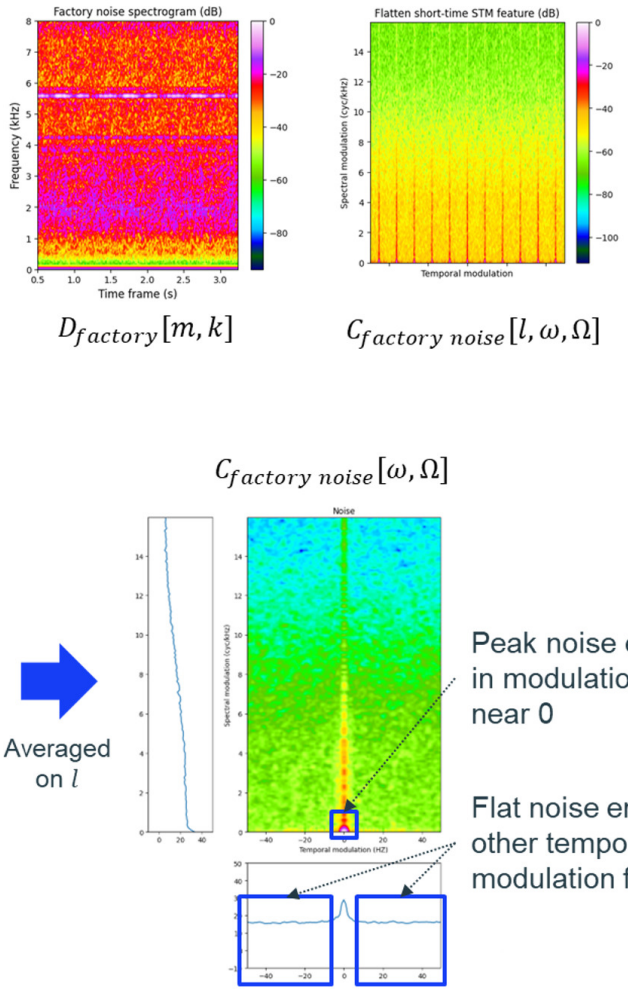


Figure 3.6: Sample of STM feature of factory noise

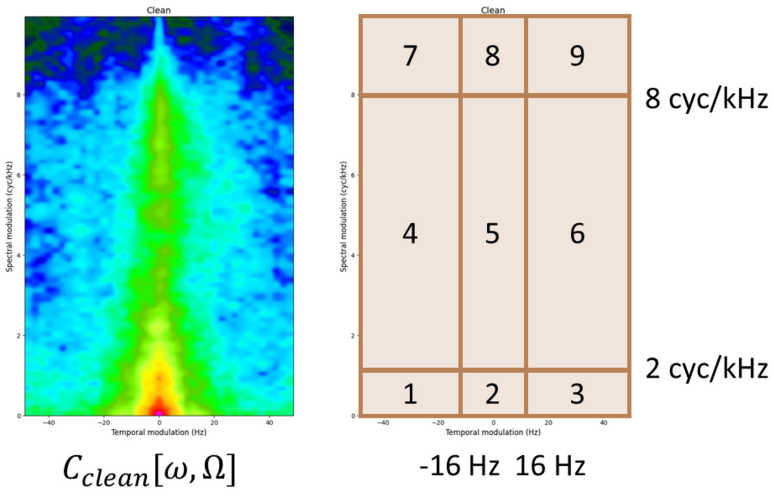


Figure 3.7: Partition of STM features into nine areas

3.3 Partitioning of the STM frequencies

As a result of the findings from the STM domain exploration, which revealed distinct energy distributions for speech and noise, a decision was made to partition the modulation frequencies into nine distinct areas. The selection of the grid lines defining these areas was informed by empirical investigations, which also aligned with the conclusions drawn in a related study [26]. In the temporal axis (horizontal axis), the grid is in $\pm$ 16 Hz, while in the spectral axis (vertical axis), it is set in 2 and 8 cyc/kHz.

This strategic division of STM frequencies allows for a more focused and targeted noise suppression, enabling a unique approach specifically designed for STM features, such as over-suppression, which can be done only to certain areas. The STM frequencies partitioning is shown in Fig. 3.7.

3.4 Noise suppression with STM and MMSE-STSA

The comprehensive AMS pipeline incorporating the STM feature is depicted in Fig. 3.8. In this proposed method, the conventional speech enhancement block is replaced with the implementation of the MMSE-STSA algorithm, utilizing the STM feature. Analogous to its spectral-domain counterpart, noise suppression solely operates on the amplitude values while retaining

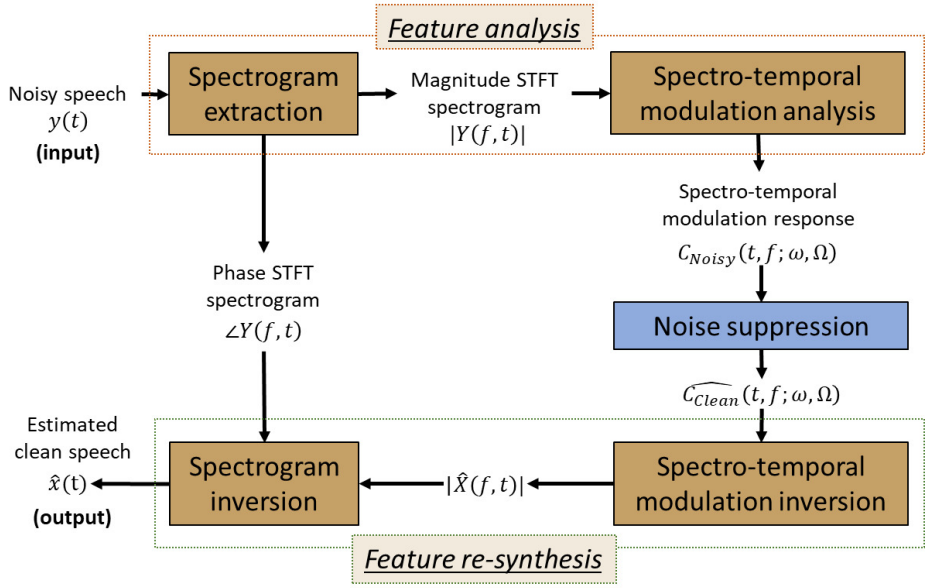


Figure 3.8: Complete AMS pipeline using STM feature

the phase information of the noisy signal to yield the reconstruction of the estimated clean signal.

This section presents a comprehensive exposition of the MMSE-STSA algorithm incorporating the STM feature. Starting with an extension of Eq. (2.20)-(2.22), the refined formulation that incorporates the STM feature to adaptively obtain the estimation of the clean speech in the presence of noise is defined.

$$\xi(\omega,\Omega) = \frac{E[|C_{clean}(\omega,\Omega)|^2]}{E[|C_{noise}(\omega,\Omega)|^2]}\,, \tag{3.3}$$

$$\gamma(\omega,\Omega) = \frac{|C_{noisy}(\omega,\Omega)|^2}{E[|C_{noise}(\omega,\Omega)|^2]}\,. \tag{3.4}$$

Using a temporary variable, $\nu(\omega,\Omega)$

$$\nu(\omega,\Omega) = \frac{\xi(\omega,\Omega)}{1+\xi(\omega,\Omega)}\gamma(\omega,\Omega)\,. \tag{3.5}$$

The gain function of the MMSE-STSA still follows the definition in Eq. (2.23).

The estimated clean spectral amplitude can be obtained by multiplying the MMSE-STSA gain with the noisy STM defined below.

$$|\hat{C_{clean}}(\omega,\Omega)| = H(\xi,\gamma)\cdot|C_{noisy}(\omega,\Omega)|\,. \tag{3.6}$$

3.4.0.1 Decision directed approach

The new estimation of the A Priori SNR can be defined as follows.

$$\begin{aligned}\hat{\xi}(\omega,\Omega) &= \frac{|C_{noisy}(\omega,\Omega)|^2 - E[|C_{noise}(\omega,\Omega)|^2]}{E[|C_{noise}(\omega,\Omega)|^2]} \\ &= \frac{|C_{noisy}(\omega,\Omega)|^2}{E[|C_{noise}(\omega,\Omega)|^2]} - 1\,.\end{aligned} \tag{3.7}$$

The estimation of the A Priori SNR obtained from the decision-directed approach can be defined as follows.

$$\begin{aligned}\hat{\xi}(l,\omega,\Omega) = \alpha &\frac{|\hat{C_{clean}}^2(l-1,\omega,\Omega)|}{E[|C_{noise}(l-1,\omega,\Omega)|^2]} \\ &+ (1-\alpha)\left(\frac{|C_{noisy}(l,\omega,\Omega)|^2}{E[|C_{noise}(l,\omega,\Omega)|^2]} - 1\right).\end{aligned} \tag{3.8}$$

3.4.0.2 Over-suppression for A Priori SNR with β

Drawing upon empirical observations, it was concluded that introducing over-suppression within the MMSE-STSA in the STM domain has the potential to enhance the noise suppression outcome. Consequently, to leverage this insight, Eq. (3.9) is adapted and modified as follows, reflecting the incorporation of over-suppression to enhance the capability to noise suppression algorithm by introducing a new variable, attenuation gain β. This adjustment aims to achieve a more refined and efficient noise suppression outcome, thereby elevating the performance and adaptability of the MMSE-STSA algorithm in the context of speech enhancement applications.

$$\hat{\xi}(\omega,\Omega) = \frac{|C_{noisy}(\omega,\Omega)|^2 - \beta\cdot E[|C_{noise}(\omega,\Omega)|^2]}{E[|C_{noise}(\omega,\Omega)|^2]}\,. \tag{3.9}$$

Hence, the redefined decision-directed approach in Eq. (3.8) is as follows:

$$\begin{aligned}\hat{\xi}(l,\omega,\Omega) = \alpha &\frac{|\hat{C_{clean}}^2(l-1,\omega,\Omega)|}{E[|C_{noise}(l-1,\omega,\Omega)|^2]} \\ &+ (1-\alpha)\left(\frac{|C_{noisy}(l,\omega,\Omega)|^2 - \beta\cdot E[|C_{noise}(l,\omega,\Omega)|^2]}{E[|C_{noise}(l,\omega,\Omega)|^2]}\right).\end{aligned} \tag{3.10}$$

The attenuation gains β may be applied to certain STM frequency groups to refine the noise suppression results. The parameter tuning for this variable is explained in detail in the following chapter.

Chapter 4

Implementation

4.1 Dataset

In this research, the clean dataset employed for experimentation originates from the Valentini dataset, a curated parallel database containing noisy and clean audio [42]. However, only the clean dataset is utilized for this study, disregarding the noisy counterparts.

The dataset contains 28 speakers, with an equal number of 14 speakers for both gender, ensuring a well-balanced representation of genders. All the selected speakers are from the same accent region, originating from England, thereby introducing a uniform linguistic characteristic.

The original data within the Valentini dataset has a uniform sampling rate of 48 kHz, ensuring a high-quality representation of the acoustic information. However, to streamline the processing and align with the specific requirements of this research, the audio is down-sampled to a 16 kHz sampling rate, striking an appropriate balance between computational efficiency and data quality while retaining valuable auditory features essential for the experimental goals of the study.

To create the noisy dataset, three types of noises are artificially additively augmented to the clean dataset. The types of noise introduced for this research include white, pink, blue, and factory noise. The diverse array of noise types is intended to simulate various real-world acoustic environments and better comprehend the performance and adaptability of the proposed noise suppression algorithm in different audio settings. Four level of SNRs are used for the experiment, which is 0, 5, 10, and 15 dB. Each SNR level represents a different scenario, ranging from a relatively low noise environment to a more challenging situation with minimal signal-to-noise separation.

From the total of 28 speakers, 4 speakers, which contain 2 male and 2 female speakers, are separated for parameter tuning. Furthermore, the remaining speakers are used for the evaluation. This is intended to fine-tune the algorithm to work well with the selected speakers and test their

performance with the remaining speakers. This approach helped to ensure that the proposed noise suppression algorithm is effective for a diverse range of speakers and can be effectively employed in real-world scenarios.

4.2 Evaluation metrics

4.2.1 Segmental SNR

Segmental SNR is a speech quality evaluation metric to evaluate the quality of a processed or degraded speech signal compared to the original clean speech [43, 44]. It is an objective measure commonly used in speech and audio processing research and engineering to quantify the distortion in degraded speech.

The idea behind segmental SNR is the segmentation of the speech waveform into shorter frames and then the calculation of the SNR for each frame. Individual segmental SNR values are then averaged to obtain a global or overall SNR score for the entire speech signal.

The original and degraded speech signals are segmented into short, adjacent frames without overlap. These frames are usually around 20-30 milliseconds in duration. For each segment, the energy of the clean referenced speech (signal energy) and the energy of the difference between the clean and degraded speech (noise energy) are computed. The segmental SNR for each segment is calculated as the ratio of the signal and noise energy in decibels using the formula defined below.

$$SegSNR(P_{signal}, P_{noise}) = 10 \cdot \log_{10} \frac{P_{signal}}{P_{noise}} . \tag{4.1}$$

The average from the individual segmental SNR values is calculated to obtain the resulting segmental SNR for the entire speech signal. This average SNR represents the global quality of the degraded speech in comparison with the reference input.

Higher Segmental SNR values indicate better speech quality, meaning that the clean speech energy dominates over the noise or distortion energy. Conversely, lower values indicate that the noise or distortion is more significant, reducing speech quality.

Segmental SNR is a simple and widely used metric for assessing speech quality, but it may not fully capture perceptual aspects of speech quality that other metrics like PESQ or STOI aim to address. Therefore, it is often combined with other metrics to provide a more comprehensive evaluation of speech-processing systems and algorithms.

4.2.2 PESQ

PESQ is a widely used speech quality assessment algorithm designed to measure the perceived speech quality in communication systems [45–47]. It was developed to evaluate telecommunications and voice-over-IP (VoIP) applications. The current usage of the PESQ is standardized by the International Telecommunication Union (ITU-T) as P.862.

The PESQ score is computed by comparing the original speech signal with the degraded version (e.g., transmitted over a network or subjected to various processing stages). The algorithm simulates the human auditory system by mimicking the human ear and brain characteristics, which are responsible for interpreting and assessing the perceived quality of speech. The quality score represents how closely the degraded address matches the original regarding perceived quality. The main components of PESQ include auditory modeling, temporal masking, perceptual weighting, comparison with a reference signal, and mapping to subjective scores.

The PESQ score is reported on a scale of -0.5 to 4.5. The upper values denote a preferred or pleasant speech quality. The general interpretation of PESQ scores can be listed as follows:

- >4.0: Excellent quality
- 3.5-4.0: Good quality
- 2.5-3.5: Fair quality
- <2.5: Poor quality

4.2.3 STOI

STOI is a speech quality measurement algorithm designed to assess the intelligibility of speech signals [48–50]. Unlike traditional speech quality metrics like PESQ, which focus on overall speech quality, STOI assesses speech intelligibility, which refers to how well the speech can be understood or comprehended by a listener.

STOI is particularly useful in scenarios where speech intelligibility is crucial, such as communication systems, hearing aid evaluation, and noise reduction algorithms. The algorithm takes the comparison of the reference signal, undistorted speech, with the noise-degraded version. STOI calculates the intelligibility score by analyzing the short-time correlation between these two signals.

Key characteristics of the STOI algorithm include short-time processing, intelligibility estimation, and perceptual weighting correlation analysis. STOI scores are typically reported on a scale of 0-1, where an STOI score close to 1 indicates a preferable speech intelligibility. In the case of an STOI

score equal to 1 means the noise-degraded speech is identical to the reference signal. Conversely, an STOI score equal to 0 denotes that the noise-degraded speech is entirely unintelligible.

STOI has become a valuable tool for researchers and engineers working on speech processing applications, especially when optimizing systems for optimal speech intelligibility in noisy or challenging environments. It complements other speech quality metrics like PESQ, providing a more comprehensive evaluation of speech communication systems.

4.2.4 LSD

Log-spectral distance (LSD) is a measure of the distance between two spectral features. The LSD is defined as follows:

$$D_{LS}(P_{signal}, P_{noise}) = \left(\frac{1}{N} \sum [\log P_{signal} - \log P_{noise}]^p \right)^{\frac{-1}{p}} , \tag{4.2}$$

where P_{signal} and P_{noise} is the power spectral of signal and noise respectively.

4.3 Parameter tuning

4.3.1 Estimation of the attenuation gain

The attenuation gain β is estimated through a grid search approach, where multiple values are tested to obtain the optimal outcome. Revisiting the partitioning of STM frequencies as depicted in Fig. 4.1. Hereafter, we refer to groups 2 and 5 as the speech-dominant groups, while the remaining groups are termed the noise-dominant groups.

Here is the step-by-step parameter tuning process.

a Change $\beta_{1,3,4,6,7,8,9}$ to some values. Then keep the best value.
b By keeping $\beta_{1,3,4,6,7,8,9}$ on the best value, change the β_5. Then keep the best value.
c By keeping β_5 to the best value, change the β_2. Then keep the best value.

The detailed result of the parameter tuning can be seen in Fig. 4.3-4.26. The best parameter is summarized in Table 4.1.

The analysis of the obtained parameter value leads to the conclusion that the optimal value of β demonstrates consistency across the selected noise types, exhibiting minimal variance. Moreover, the attenuation gain

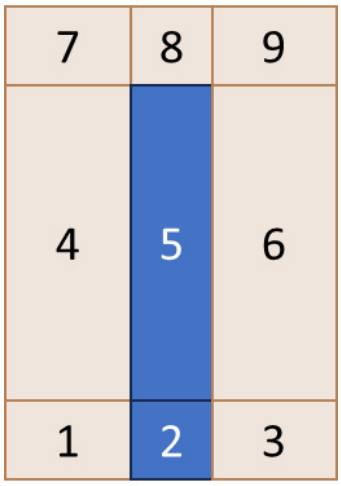


Figure 4.1: STM frequencies partitioning. Blue is the speech-dominant area; brown is the noise-dominant area.

Table 4.1: Summary of the tuning of the attenuation gain β

Noise type	$\beta_{1,3,4,6,7,8,9}$	β_5	β_2
White	5	2	1
Pink	5	2	1
Factory	5	2	2

remains relatively stable SNR level changes. Notably, selecting larger β values for noise-dominant regions yields enhanced noise suppression results. The resulting β value, obtained through this exploration, is subsequently employed for evaluation purposes, utilizing the test set to conduct the assessment of the proposed noise suppression method.

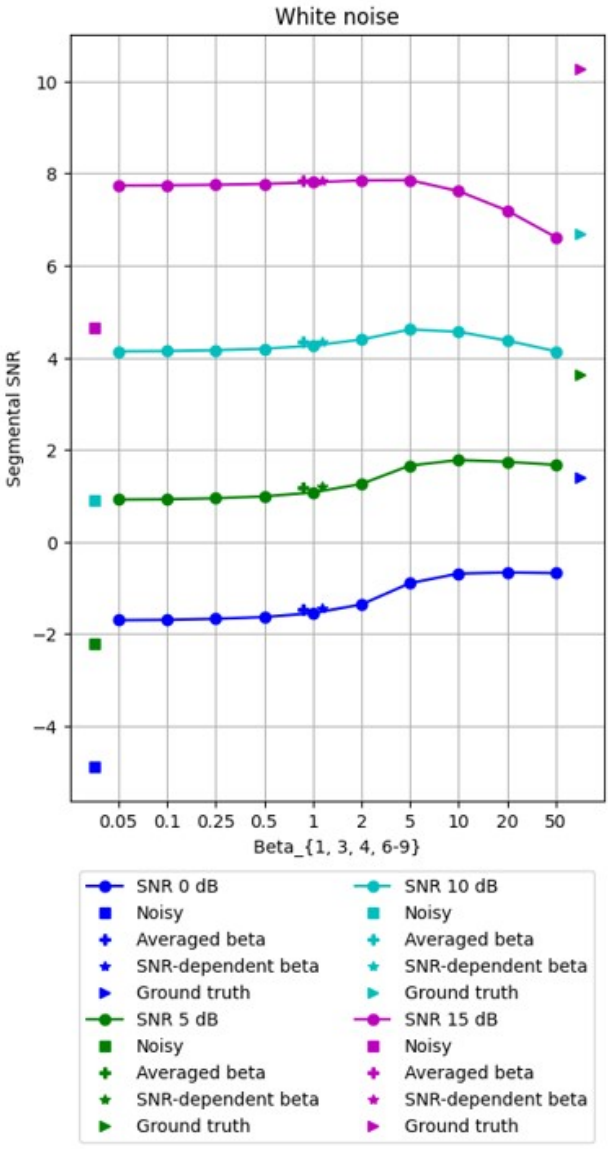


Figure 4.2: Tuning the $\beta_{1,3,4,6,7,8,9}$. White noise. Segmental SNR score

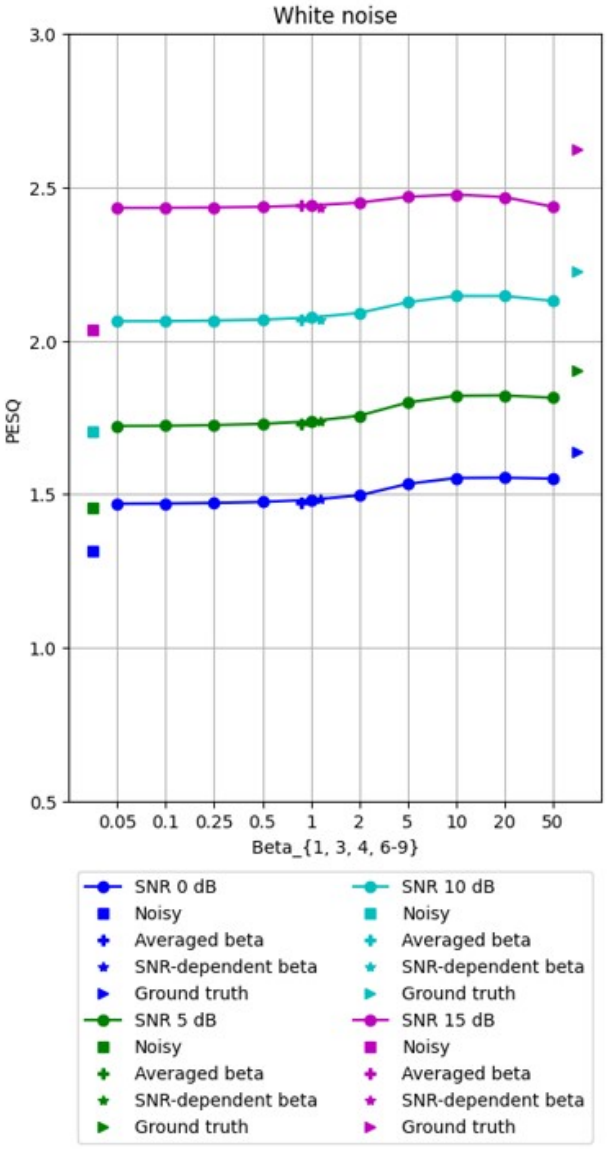


Figure 4.3: Tuning the $\beta_{1,3,4,6,7,8,9}$. White noise. PESQ score

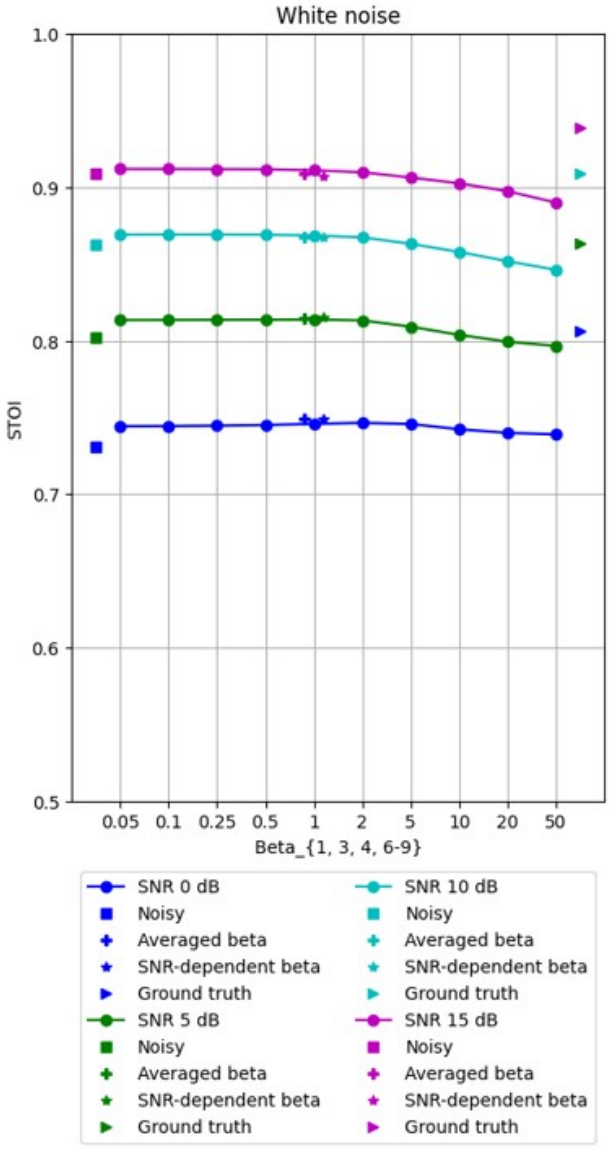


Figure 4.4: Tuning the $\beta_{1,3,4,6,7,8,9}$. White noise. STOI score

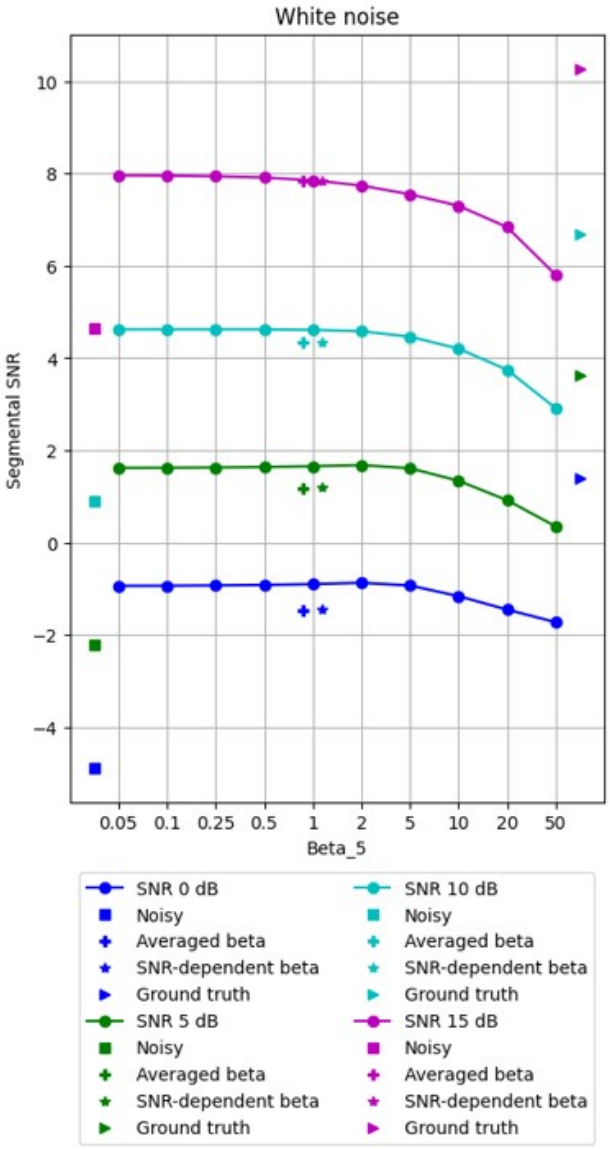


Figure 4.5: Tuning the β_5. White noise. Segmental SNR score

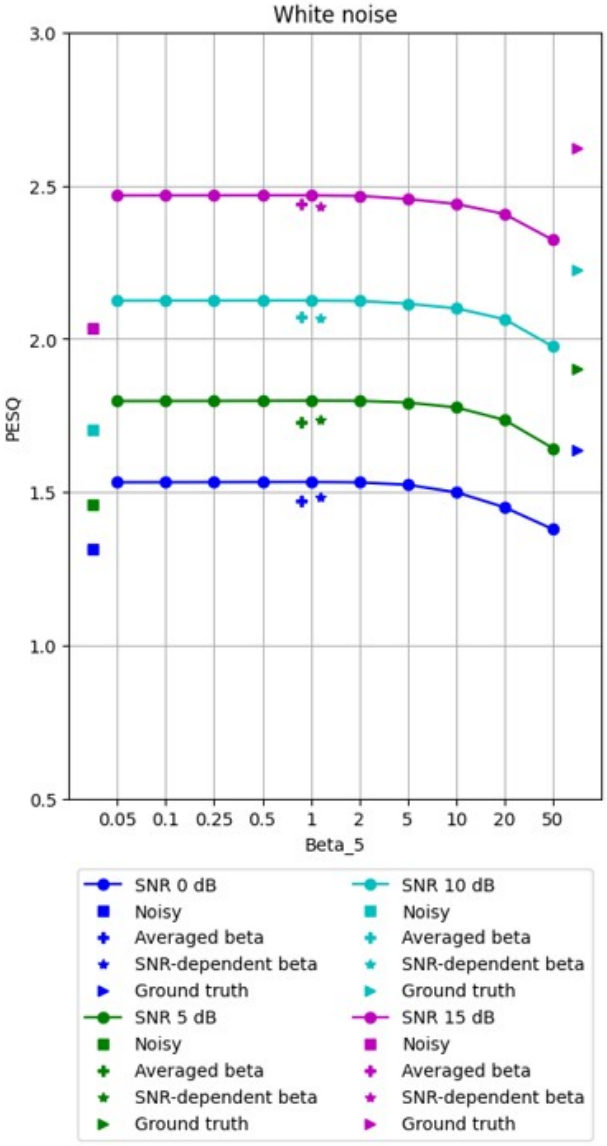


Figure 4.6: Tuning the β_5. White noise. PESQ score

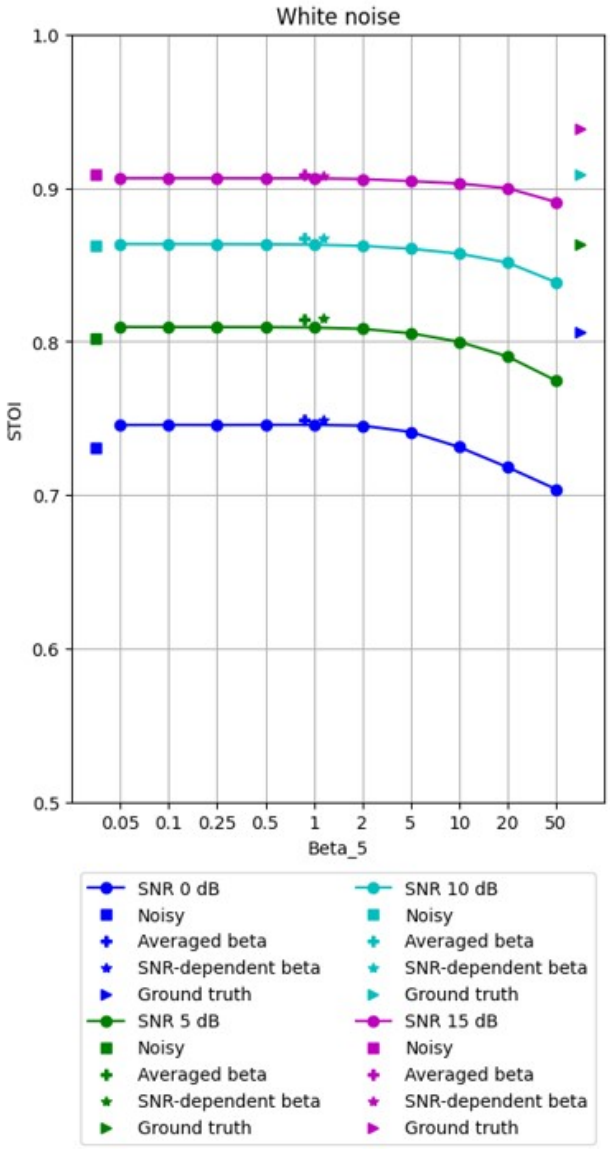


Figure 4.7: Tuning the β_5. White noise. STOI score

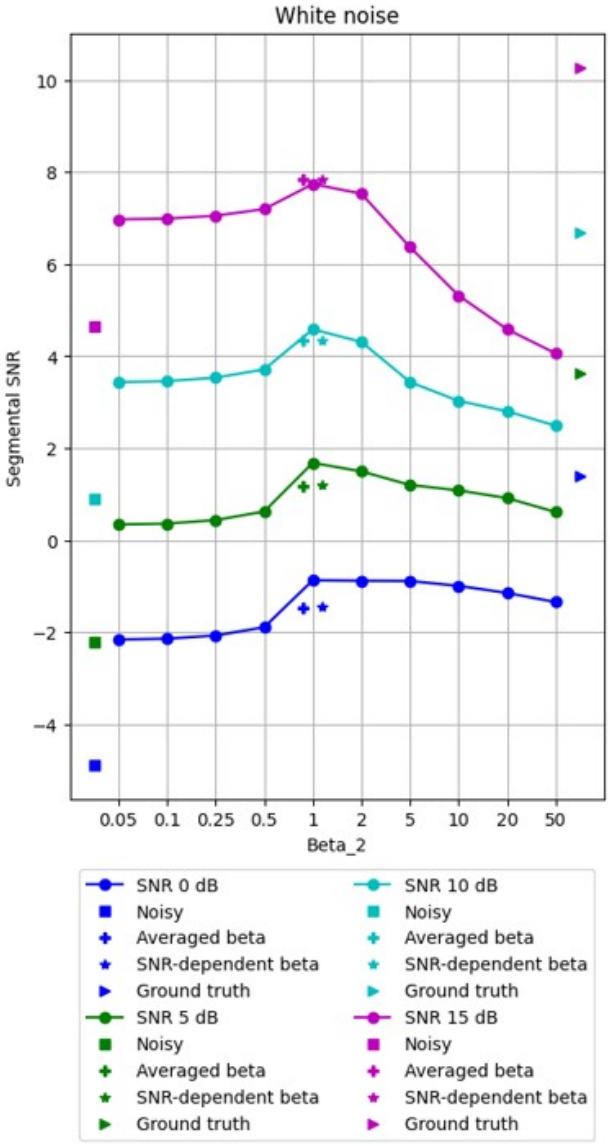


Figure 4.8: Tuning the β_2. White noise. Segmental SNR score

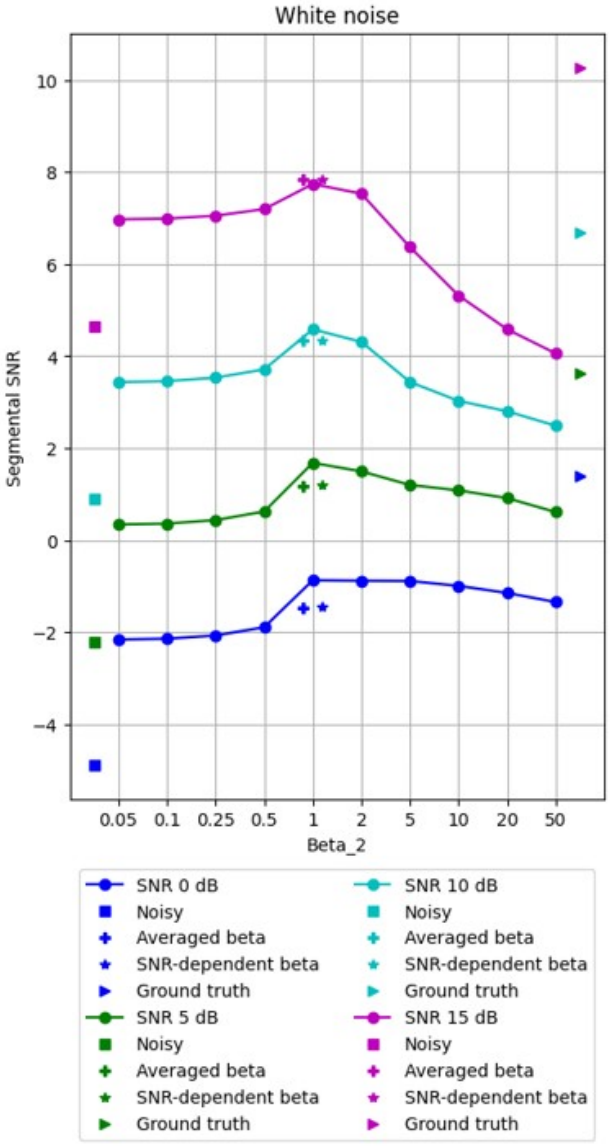


Figure 4.9: Tuning the β_2. White noise. PESQ score

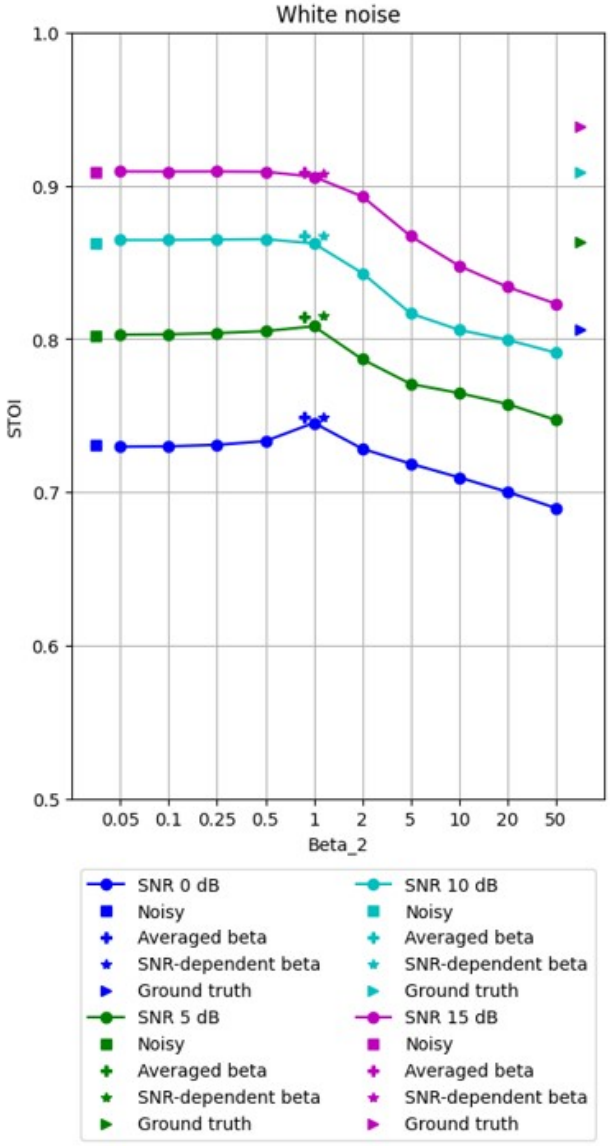


Figure 4.10: Tuning the β_2. White noise. STOI score

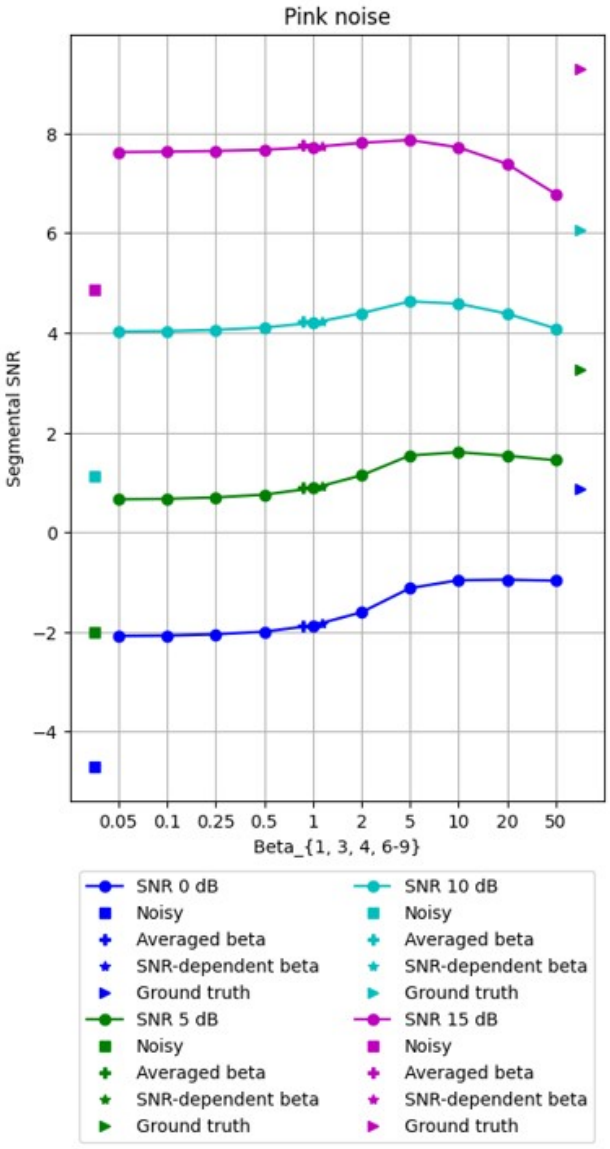


Figure 4.11: Tuning the $\beta_{1,3,4,6,7,8,9}$. Pink noise. Segmental SNR score

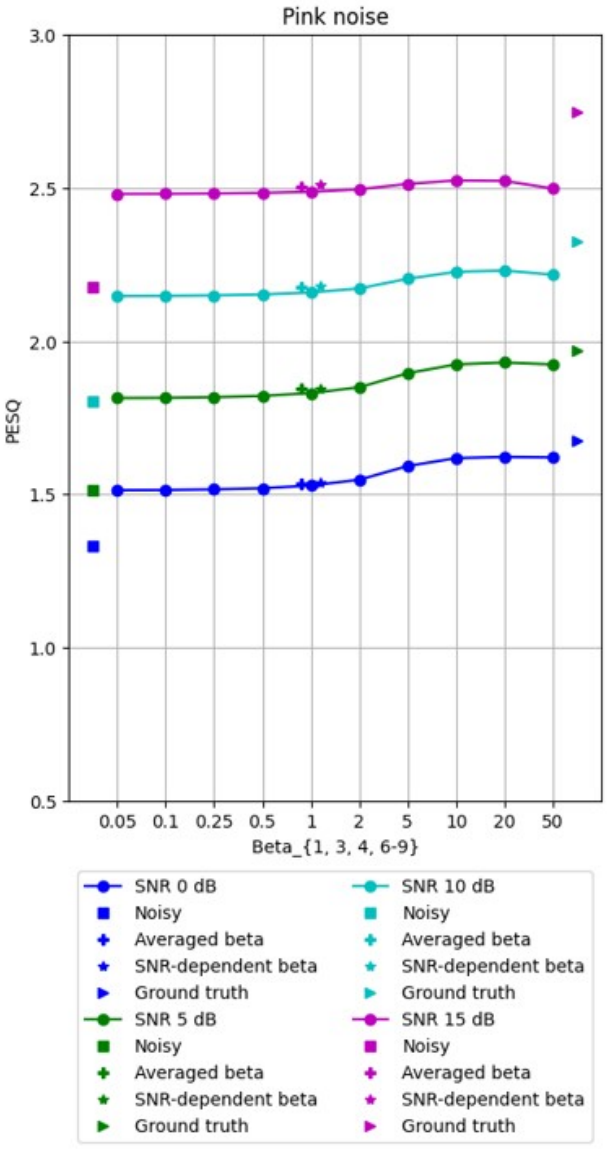


Figure 4.12: Tuning the $\beta_{1,3,4,6,7,8,9}$. Pink noise. PESQ score

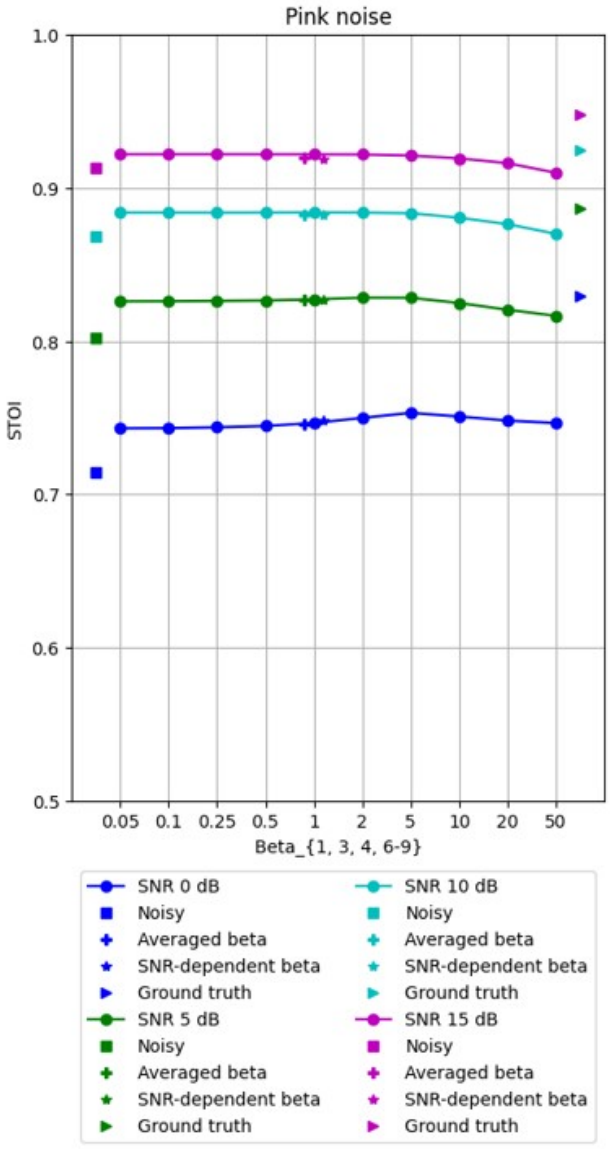


Figure 4.13: Tuning the $\beta_{1,3,4,6,7,8,9}$. Pink noise. STOI score

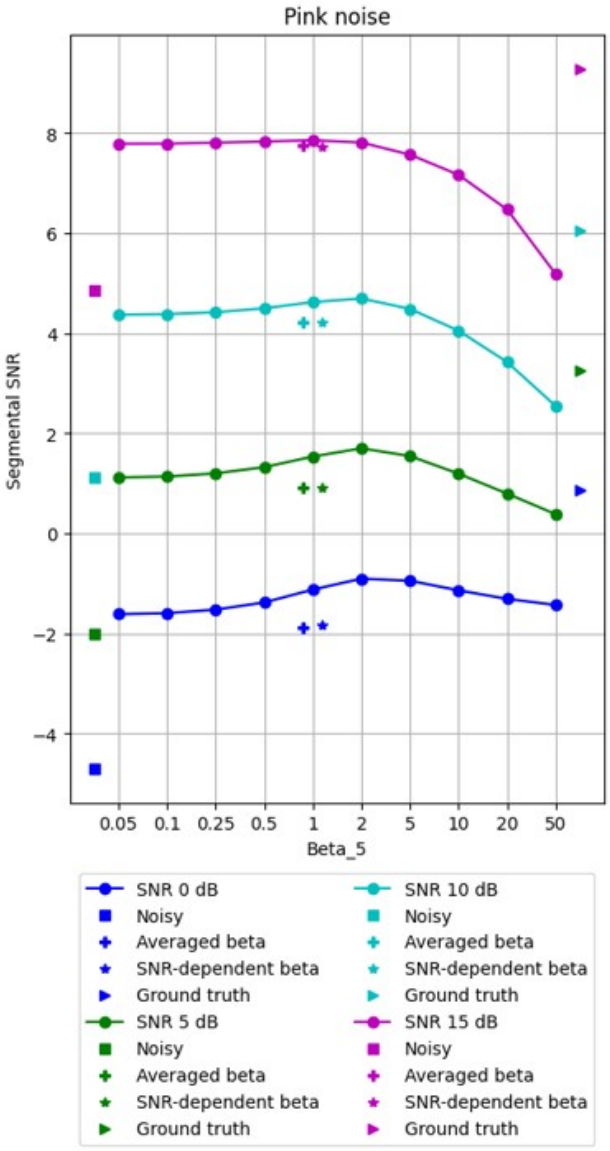


Figure 4.14: Tuning the β_5. Pink noise. Segmental SNR score

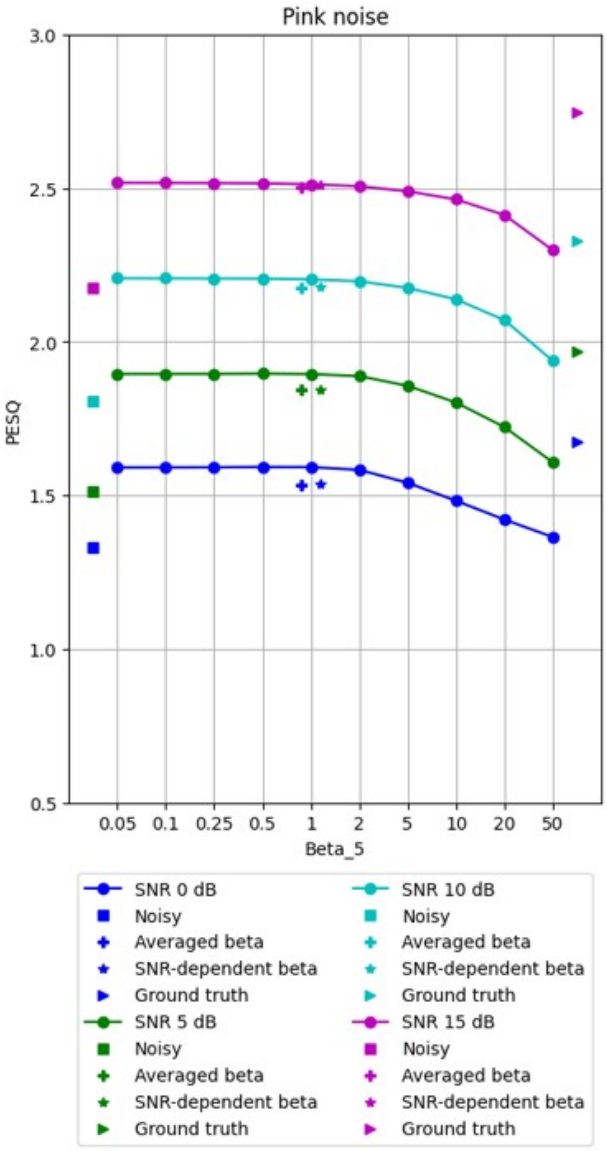


Figure 4.15: Tuning the β_5. Pink noise. PESQ score

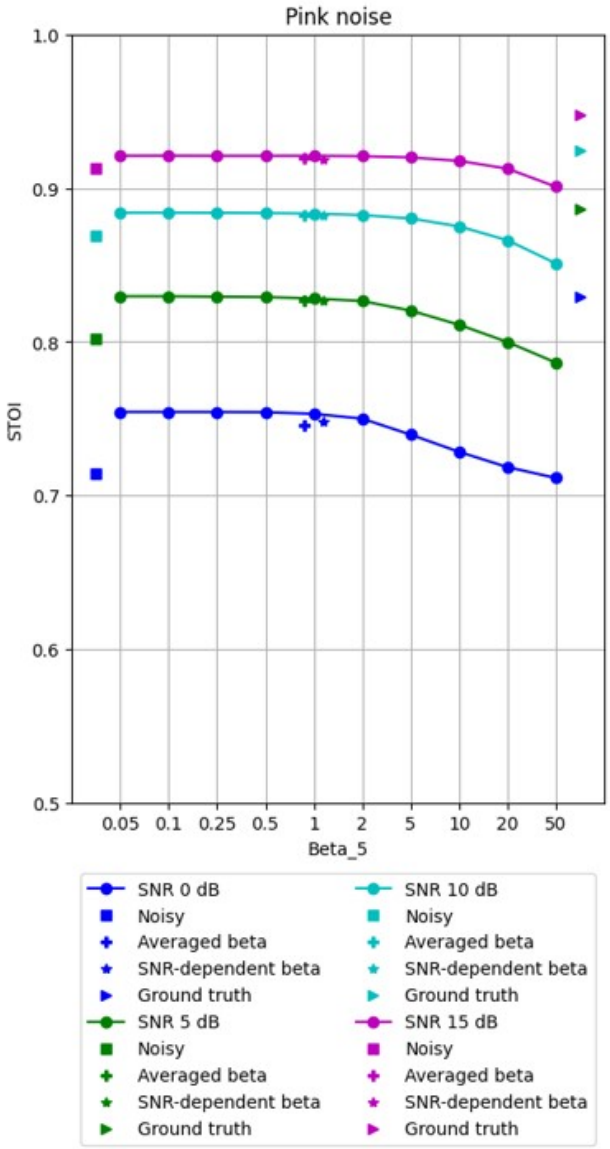


Figure 4.16: Tuning the β_5. Pink noise. STOI score

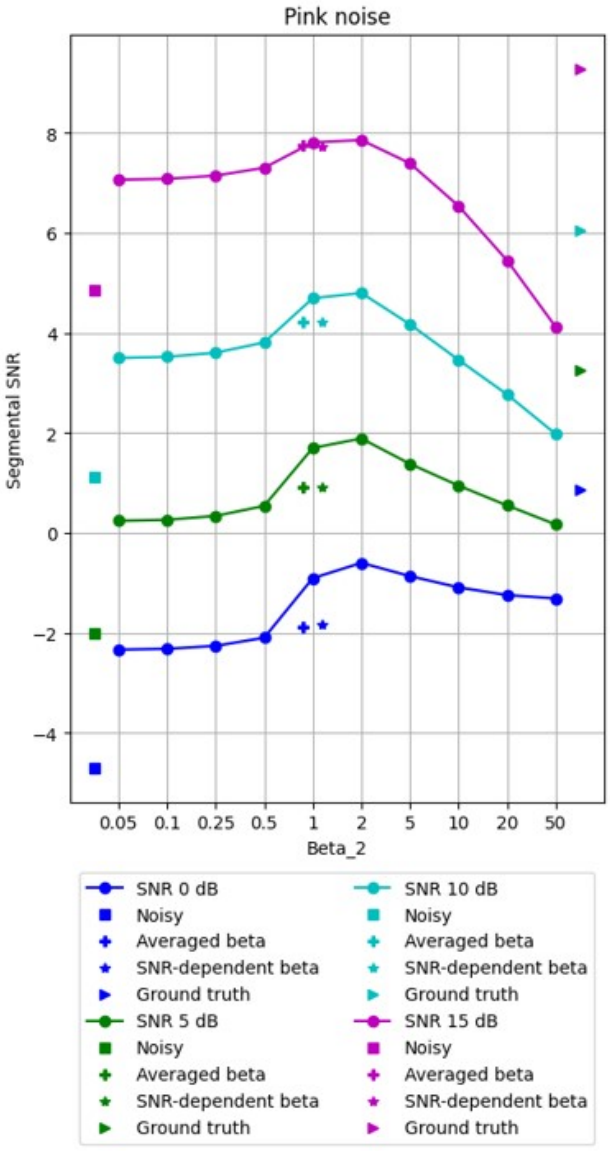


Figure 4.17: Tuning the β_2. Pink noise. Segmental SNR score

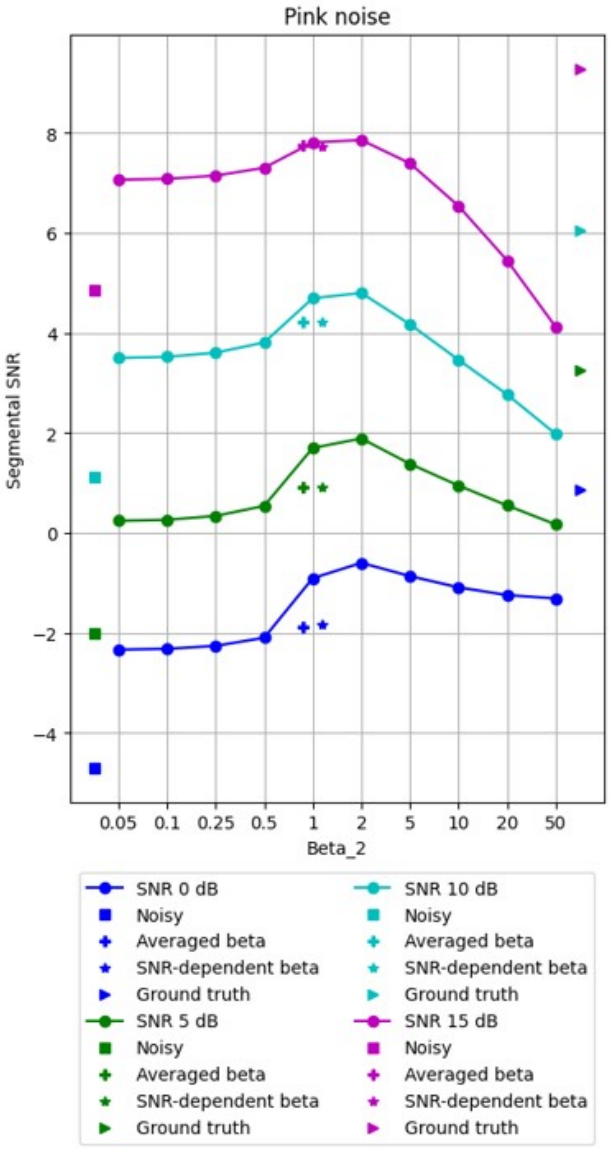


Figure 4.18: Tuning the β_2. Pink noise. PESQ score

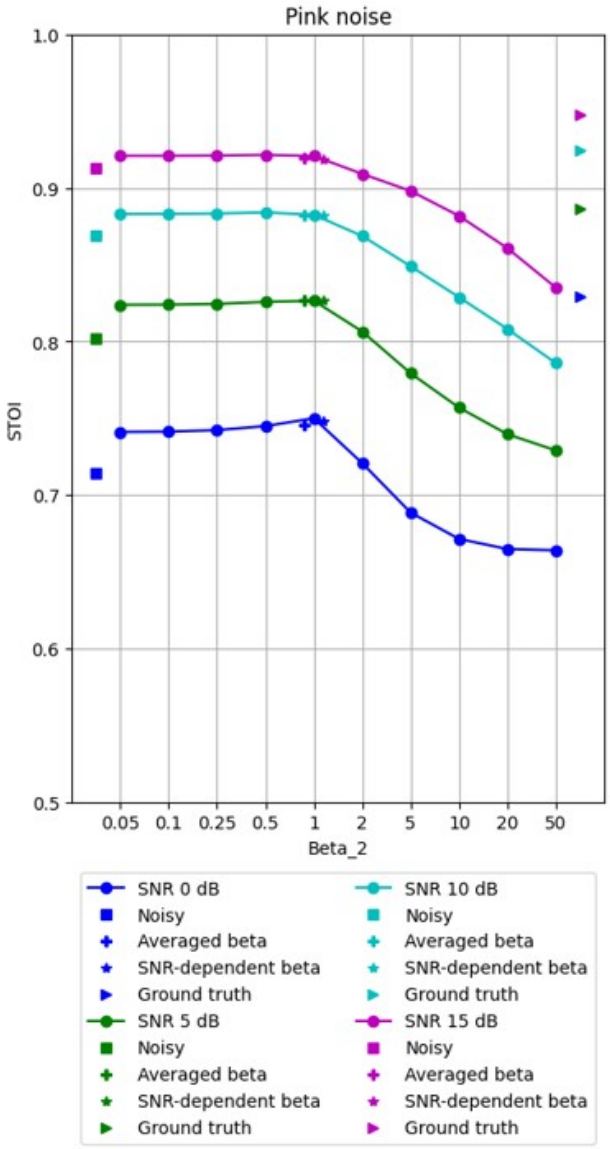


Figure 4.19: Tuning the β_2. Pink noise. STOI score

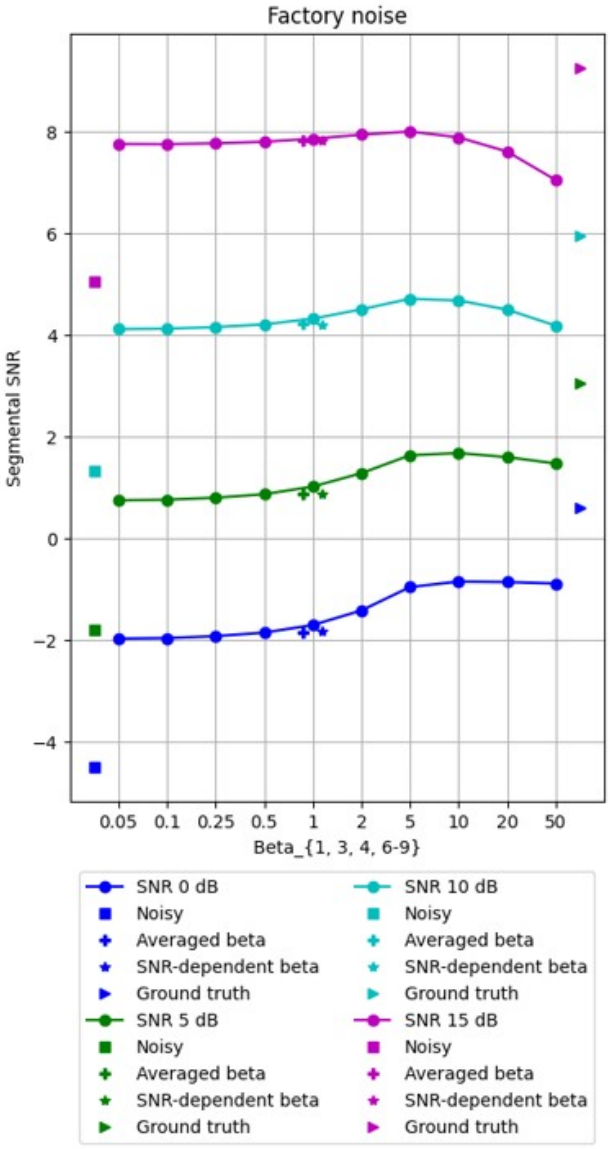


Figure 4.20: Tuning the $\beta_{1,3,4,6,7,8,9}$. Factory noise. Segmental SNR score

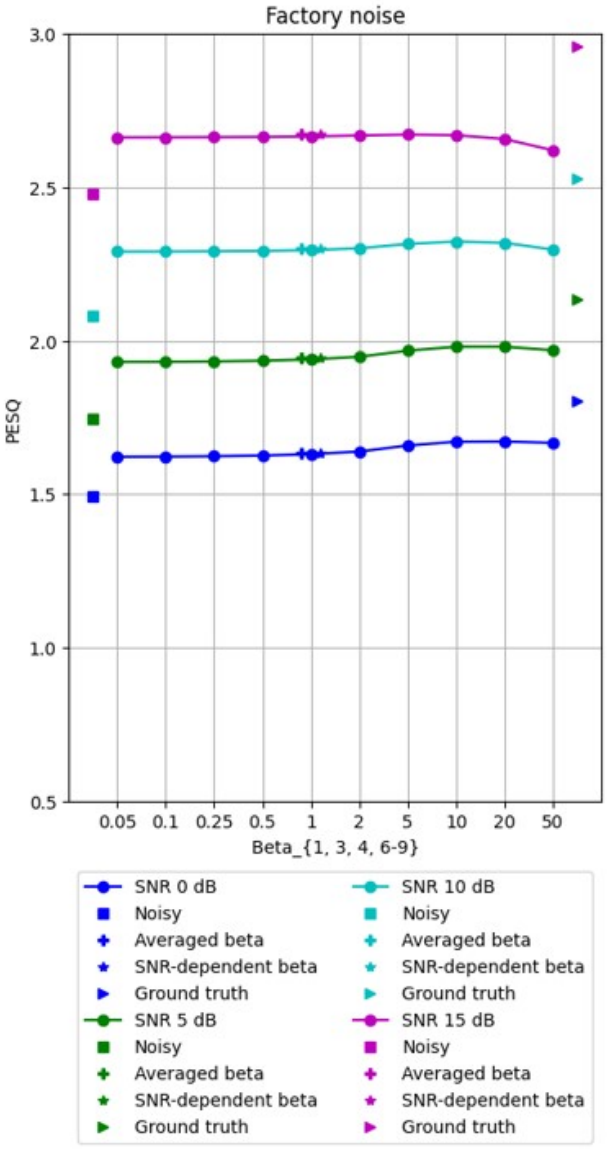


Figure 4.21: Tuning the $\beta_{1,3,4,6,7,8,9}$. Factory noise. PESQ score

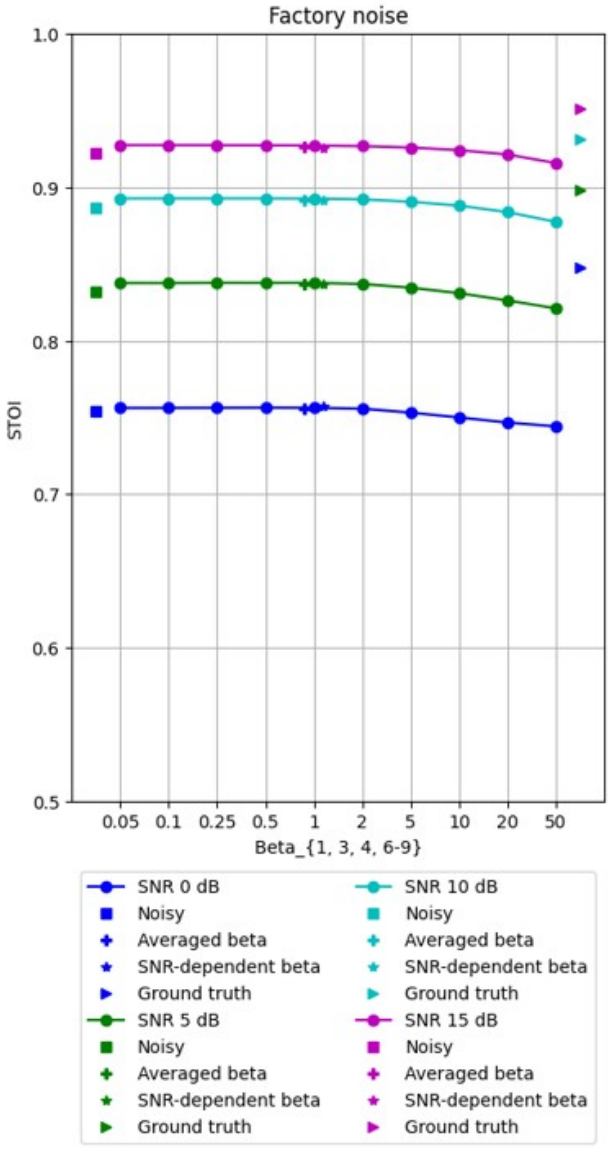


Figure 4.22: Tuning the $\beta_{1,3,4,6,7,8,9}$. Factory noise. STOI score

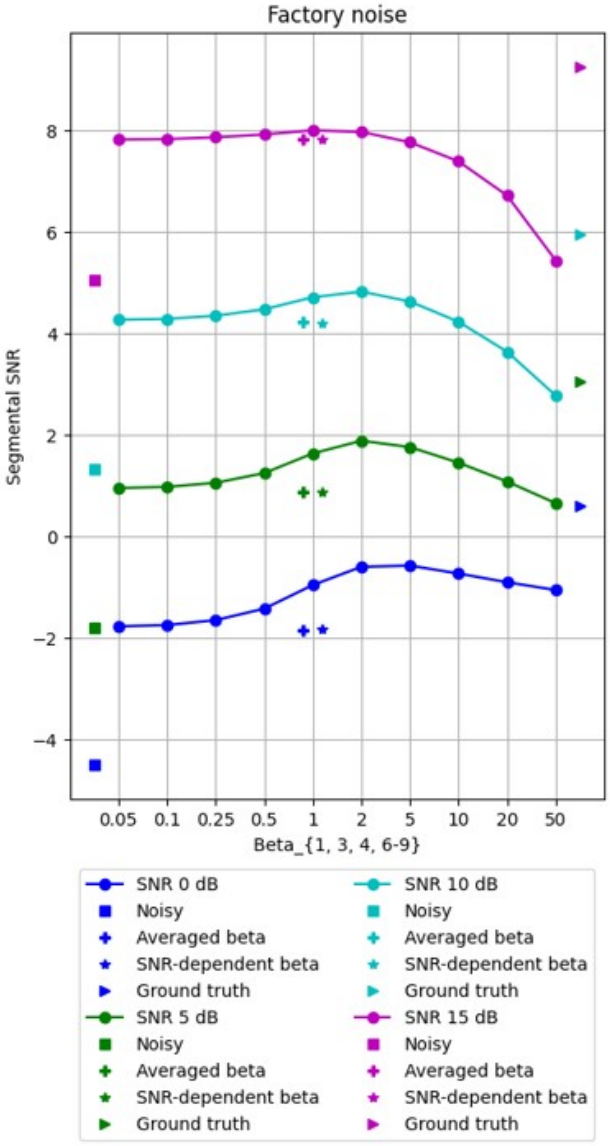


Figure 4.23: Tuning the β_5. Factory noise. Segmental SNR score

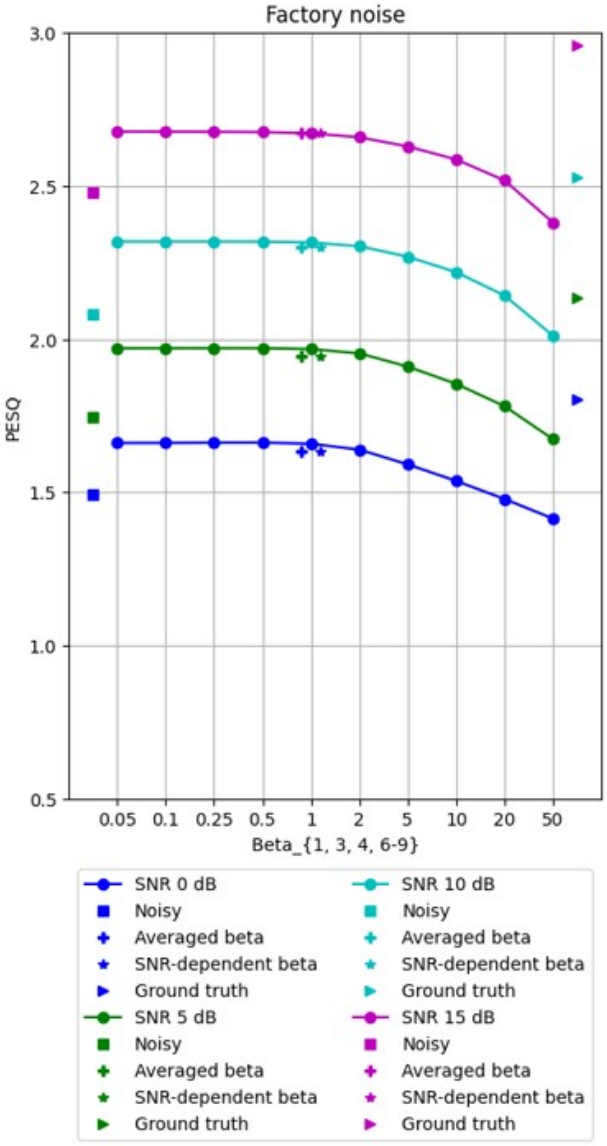


Figure 4.24: Tuning the β_5. Factory noise. PESQ score

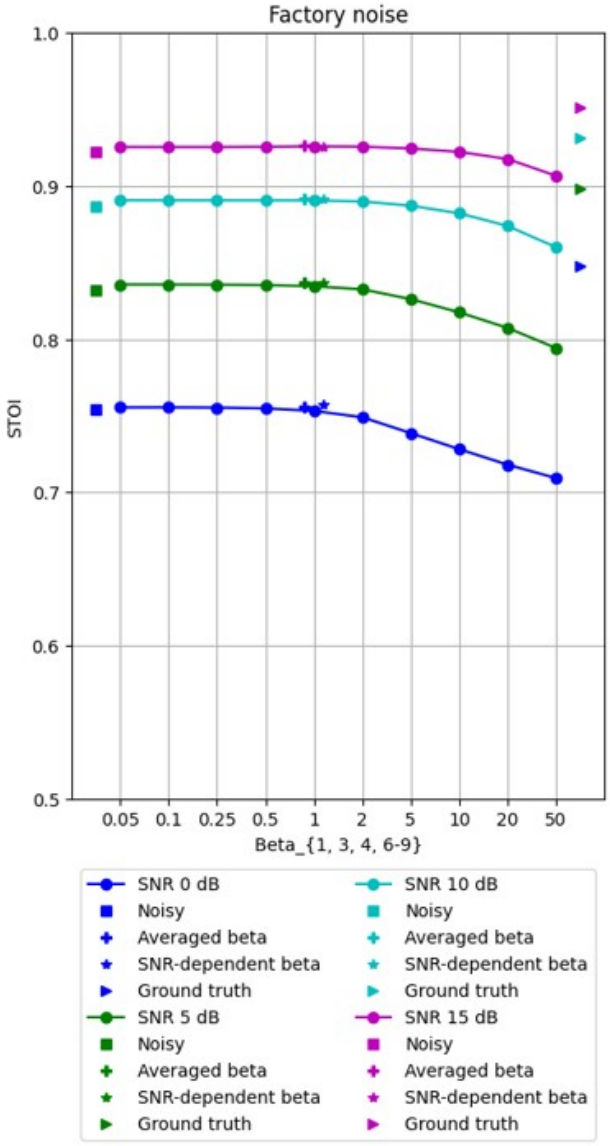


Figure 4.25: Tuning the β_5. Factory noise. STOI score

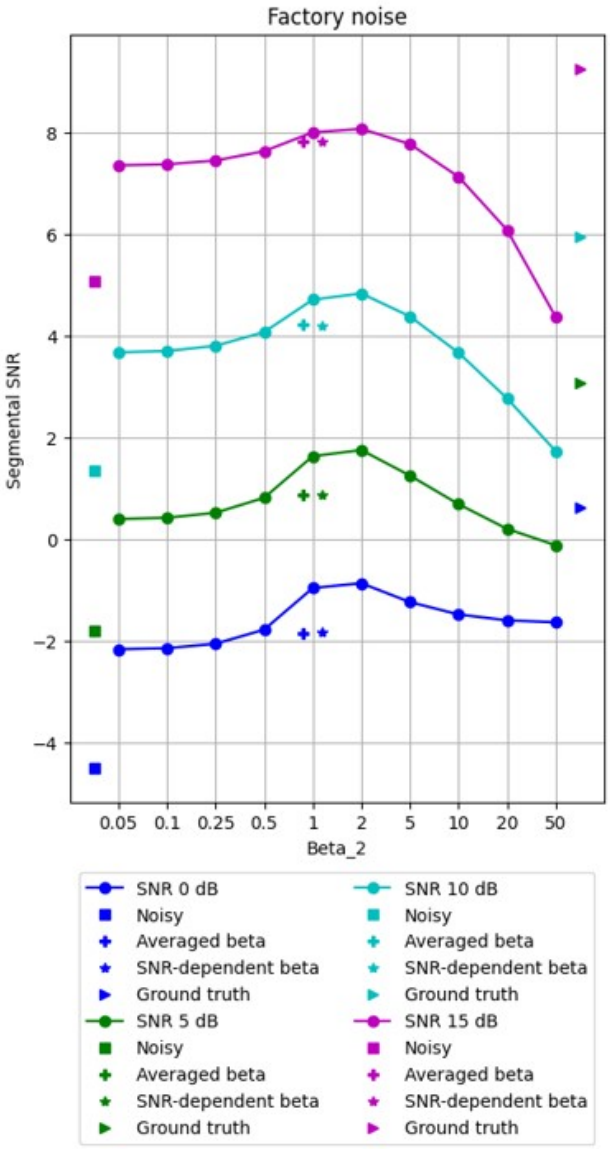


Figure 4.26: Tuning the β_2. Factory noise. Segmental SNR score

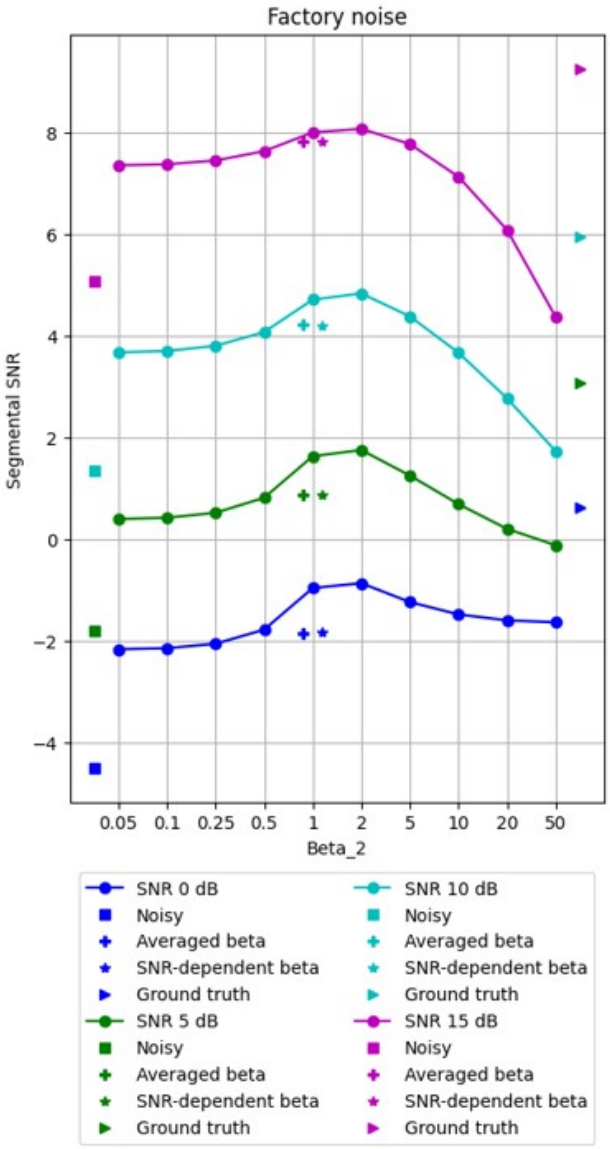


Figure 4.27: Tuning the β_2. Factory noise. PESQ score

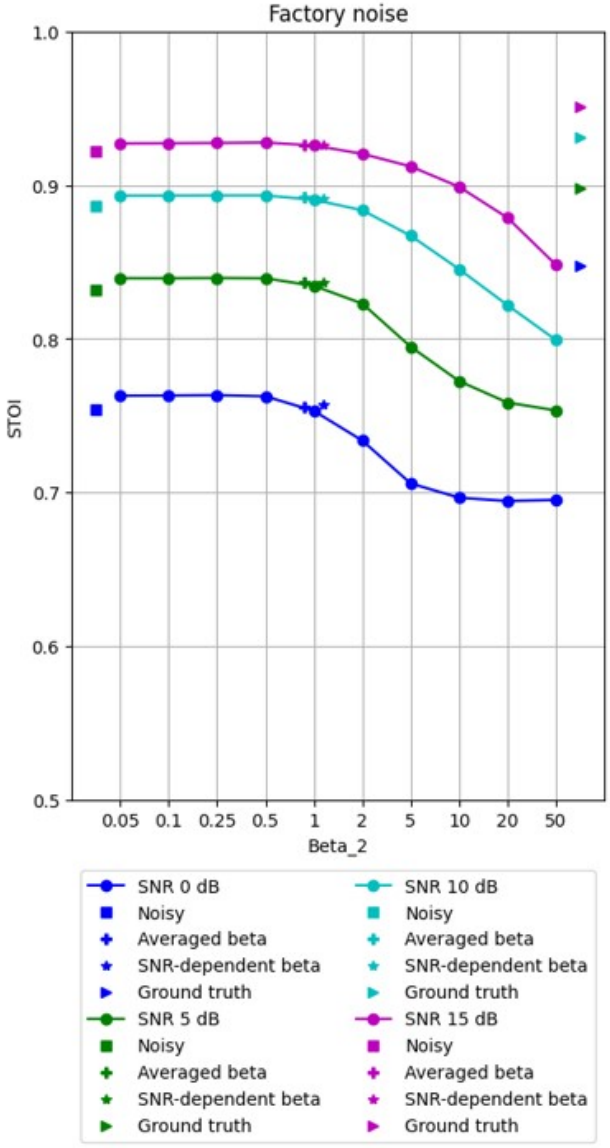


Figure 4.28: Tuning the β_2. Factory noise. STOI score

Chapter 5

Evaluation

5.1 Evaluation of the AMS framework using STM feature

This section aims to validate the correctness and accuracy of the proposed AMS framework for the STM feature. Additionally, it measures the ground-truth value where the modification step is not applied. The evaluation of the ASM framework was conducted using the same dataset utilized for parameter tuning.

It is important to highlight that in the AMS framework for the STM feature, the noisy phase information is employed for the inversion step during both spectrogram and signal reconstruction. However, for additional insights, results using the clean STM phase information for reconstructing the spectrogram are also presented to assess the significance of the STM phase information.

The evaluation is conducted by measuring the log spectral distance (LSD) from the reference and the estimated spectrogram. Three evaluation metrics are also utilized to measure the quality and the intelligibility of the reconstructed signal.

The result in Table 5.1 is for the noisy speech augmented by adding white noise. In this AMS framework, the normal noisy STM phase information is used during the reconstruction process. The result in this table can be regarded as the ground-truth value for noise suppression using the joint STM feature. The result in Table 5.2 represents the evaluation result of the AMS framework while the clean STM phase information is utilized. The phase information for the spectrogram is always controlled to be the noisy phase information.

The very small value of the LSD score in Table 5.2 indicates the perfect reconstruction from the STM feature back to the spectrogram using both clean amplitude and phase information. In contrast, a considerable difference is observed in the LSD score in Table 5.1 when the STM phase information is substituted with the one from the noisy signal.

Table 5.1: Evaluation of AMS framework using *noisy* STM phase information

SNR	$LSD(X, \hat{X})$	$SegSNR(x, \hat{x})$	$PESQ(x, \hat{x})$	$STOI(x, \hat{x})$
0	1.94 (± 0.24)	0.53 (± 0.37)	1.80 (±0.14)	0.91 (± 0.001)
5	1.69 (± 0.25)	2.90 (± 0.28)	2.11 (± 0.21)	0.95 (± 0.005)
10	1.41 (± 0.23)	6.04 (± 0.17)	2.47 (± 0.22)	0.98 (± 0.003)
15	1.13 (± 0.16)	9.58 (± 0.18)	2.94 (± 0.14)	0.99 (± 0.002)

Table 5.2: Evaluation of AMS framework using *clean* STM phase information

SNR	$LSD(X, \hat{X})$	$SegSNR(x, \hat{x})$	$PESQ(x, \hat{x})$	$STOI(x, \hat{x})$
0	5.04e-06 (± 9.13e-07)	3.06 (± 0.26)	3.87 (± 0.49)	0.98 (± 0.009)
5	5.05e-06 (± 9.13e-07)	5.22 (± 0.38)	3.93 (± 0.04)	0.99 (± 0.002)
10	5.05e-06 (± 9.13e-07)	7.57 (± 0.43)	4.14 (± 0.06)	0.99 (± 0.001)
15	5.05e-06 (± 9.13e-07)	10.12 (± 0.52)	4.31 (± 0.01)	0.99 (± 0.001)

Measuring the differences between the segmental SNR, PESQ, and STOI scores, indicates the importance of the phase information for the STM feature.

5.2 Results

The evaluation result for noise suppression using the STM feature and MMSE-STSA is shown in Figs. 5.1 to 5.3. The horizontal axis represents the SNR levels of 0-15 dB, while the vertical axis is the evaluation score with the respective evaluation metrics.

The evaluation results using segmental SNR are presented in three figures, with Fig. 5.1 for white noise, Fig. 5.2 for the pink noise, and Fig. 5.3 is for the factory noise. Consequently, the evaluation results using PESQ are in Figs. 5.4 to 5.6 and STOI are in Figs. 5.7 to 5.9.

5.3 Discussion

The results obtained from the evaluation using the STM technique are depicted in Figure 5.10. In accordance with these results, it is observed that the noise reduction method employing STM and MMSE demonstrates an effective reduction in noise levels, albeit with the concurrent occurrence of over-suppression.

This initial evaluation serves as a foundation for the further enhancement of the noise reduction process. To achieve this, strategies for optimizing the observed over-suppression, while simultaneously maintaining a reduction in noise levels, need to be devised. Through the systematic adjustment of parameters and operational settings inherent to the STM and MMSE-based approach, experiments are undertaken to cultivate a balance between noise reduction and an over-suppression algorithm. It is crucial to underscore that this process of refinement is informed by empirical insights. In this regard, our research is underscored by a dedication to methodological optimization and the pursuit of acoustic excellence in challenging environments.

The noise suppression result of the proposed method is presented using three evaluation metrics, segmental SNR, PESQ, and STOI. To benchmark the efficacy of the MMSE-STSA algorithm using the STM feature, a comparison with two other statistical algorithms was made, namely with the Wiener filter and the MMSE-STSA algorithm, which both utilize spectral features.

Segmental SNR reflects the level of distortion between the noise-suppressed and the clean audio. As shown in Figs. 5.1 to 5.3, the proposed method demonstrates an improvement in distortion compared to noisy speech, though the score is not the highest achieved. The noise suppression algorithm proposed in this study is on par with the Wiener filter. However, the original MMSE-STSA algorithm, operating in the spectral domain, yields the best overall result across noise types. Importantly, it is worth noting that the ground-truth score obtained using the STM feature obtains a lower or similar score in segmental SNR, particularly in lower SNR levels (i.e., 0 and 5 dB), as it is compared to the original MMSE-STSA algorithm.

For reference, the ground-truth result is obtained by substituting the noisy amplitude with the clean amplitude in the STM domain. During the synthesis step, the noisy phase is used for both synthesizing the spectrogram and the signal. This ensures that the MMSE-STSA using the STM features considers the characteristics of the noisy signal during the synthesis process, contributing to the overall noise suppression performance. However, the comparison with the original MMSE-STSA algorithm indicates that the proposed method may exhibit limitedness in achieving the highest segmental

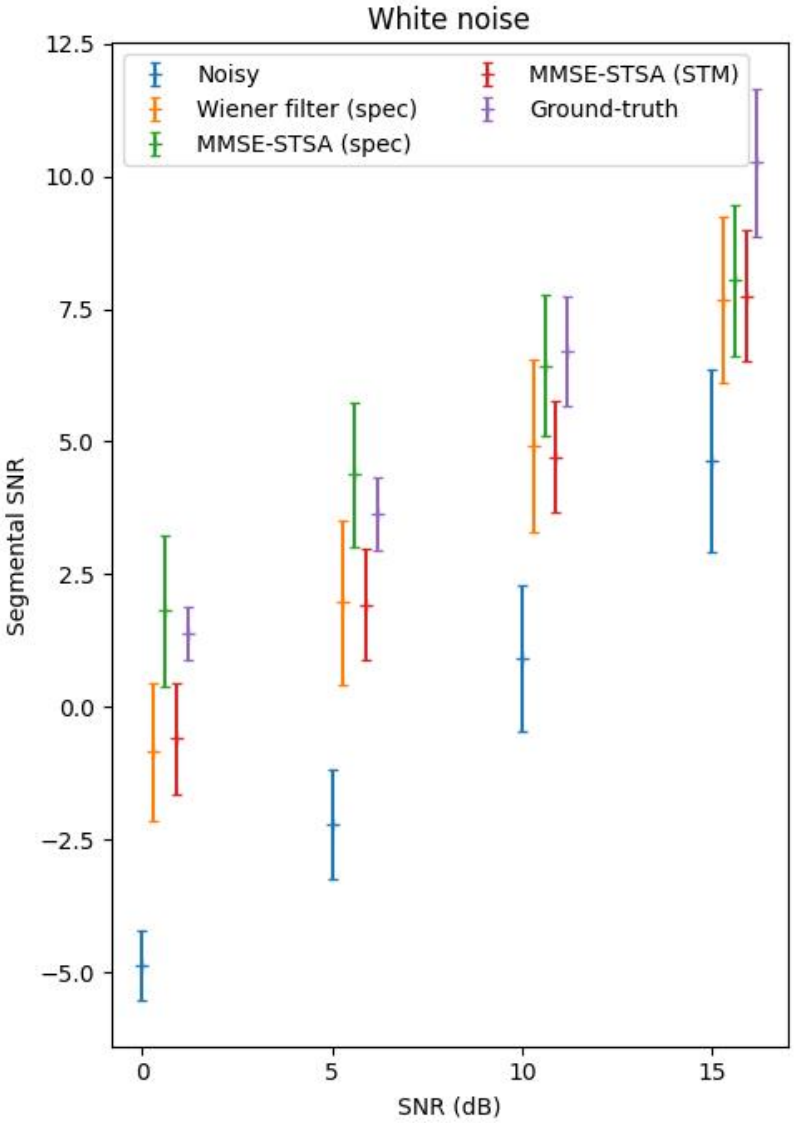


Figure 5.1: Segmental SNR: white noise

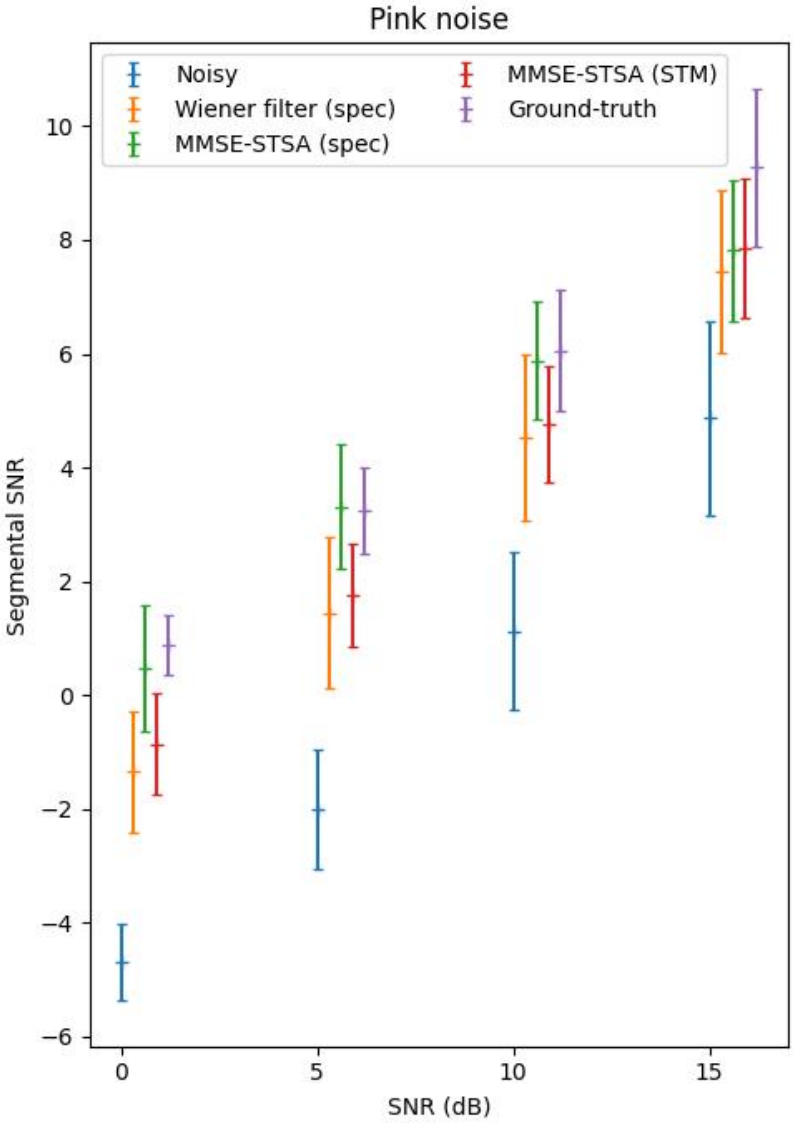


Figure 5.2: Segmental SNR: pink noise

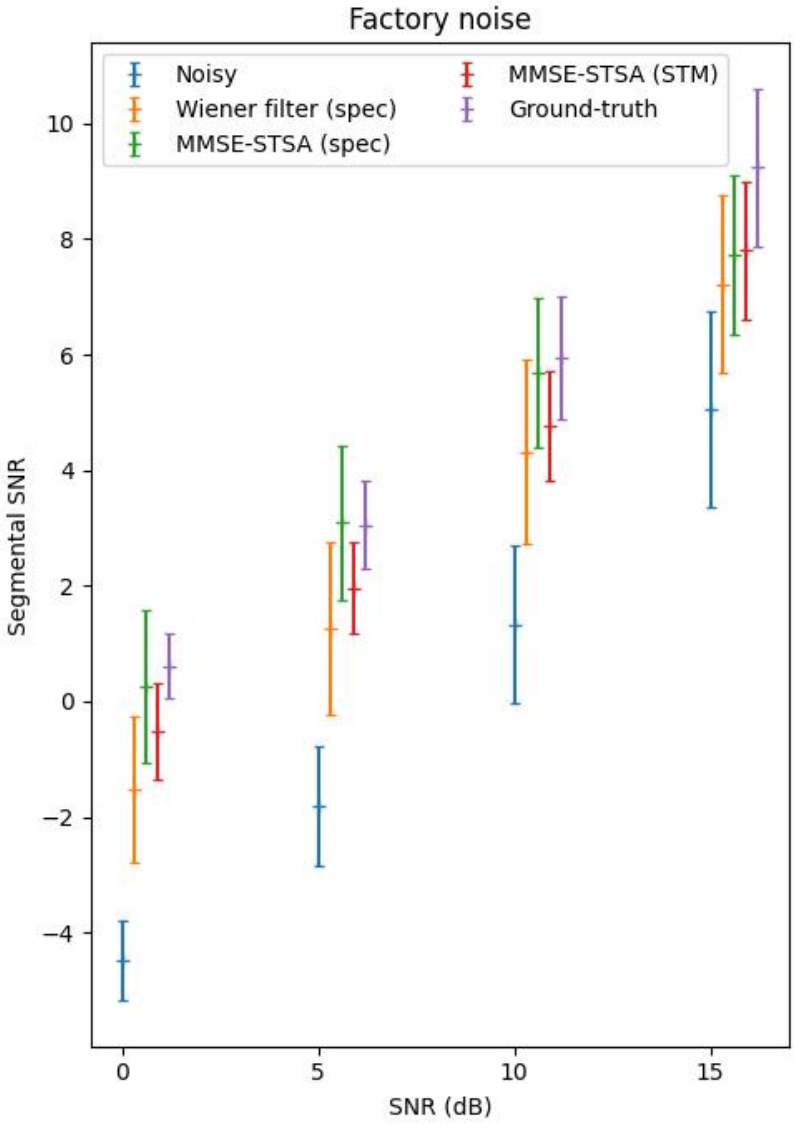


Figure 5.3: Segmental SNR: factory noise

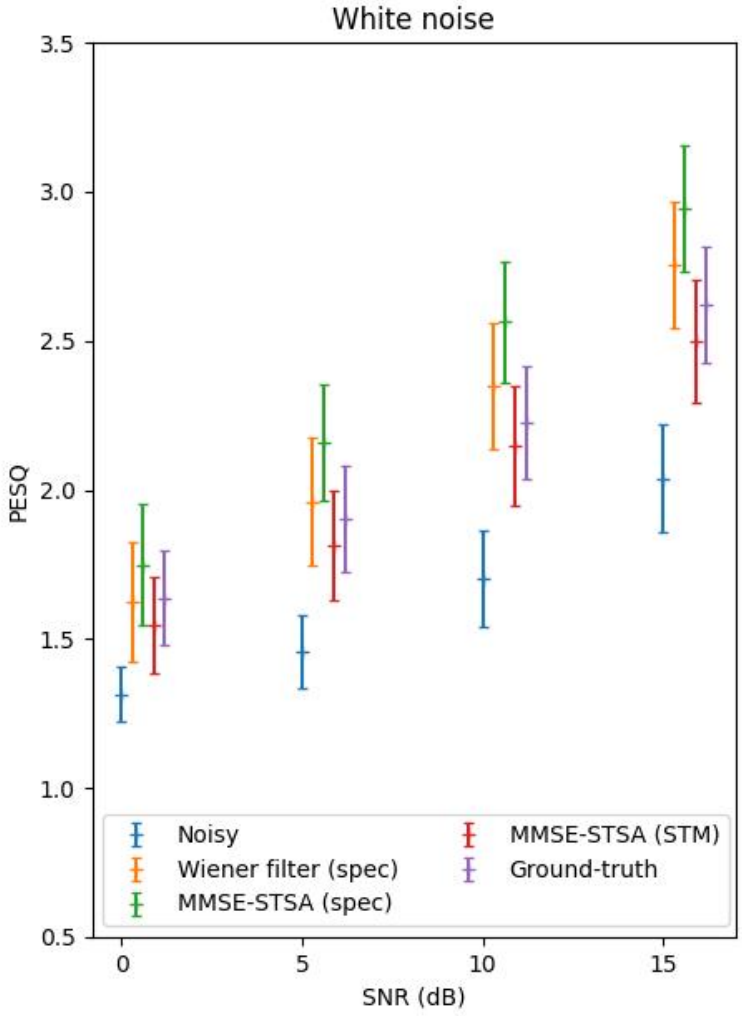


Figure 5.4: PESQ score: white noise

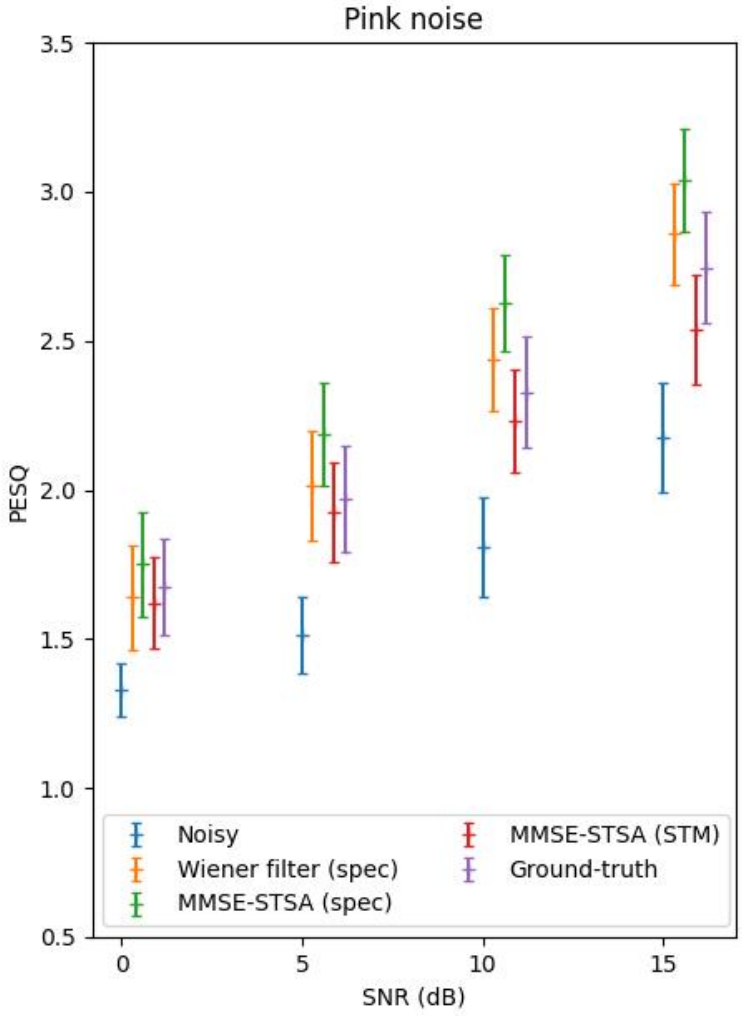


Figure 5.5: PESQ score: pink noise

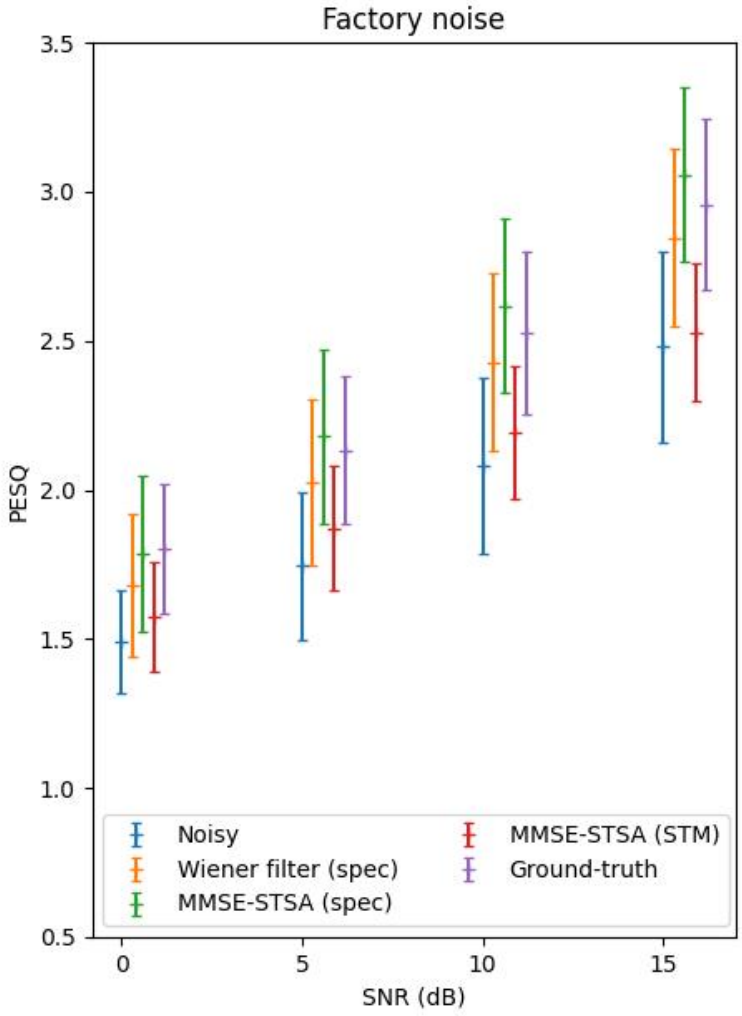


Figure 5.6: PESQ score: factory noise

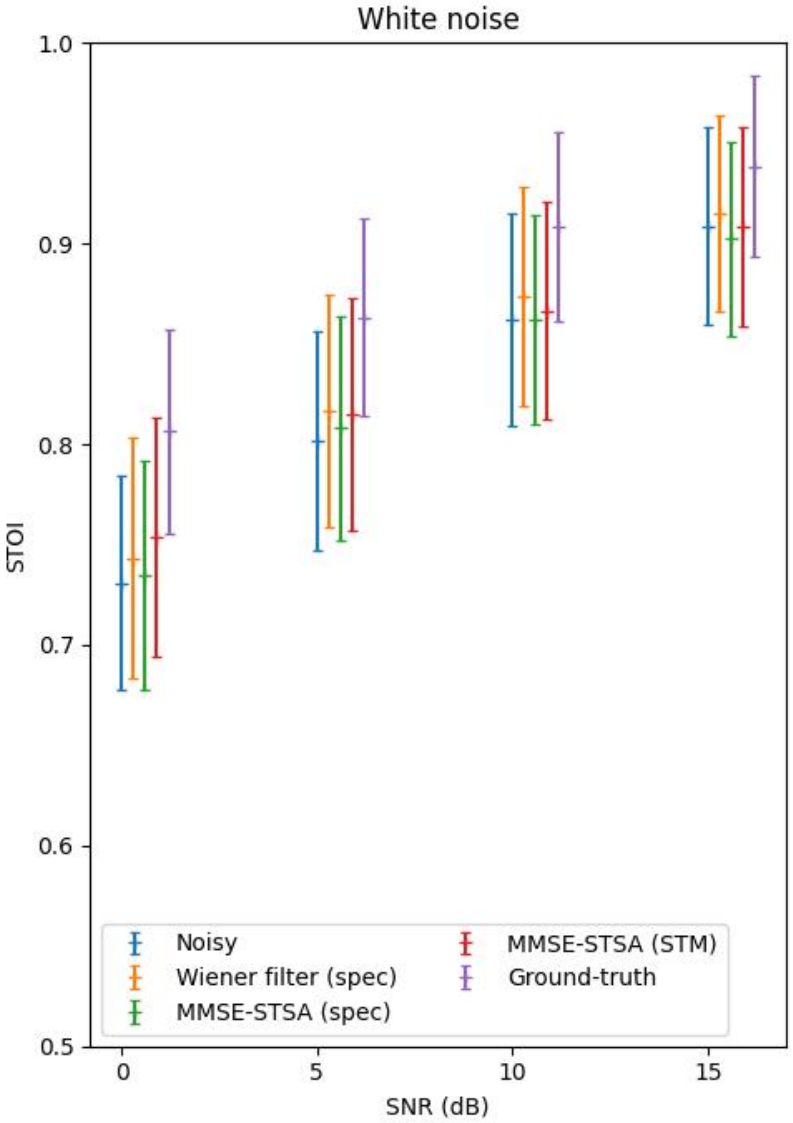


Figure 5.7: STOI score: white noise

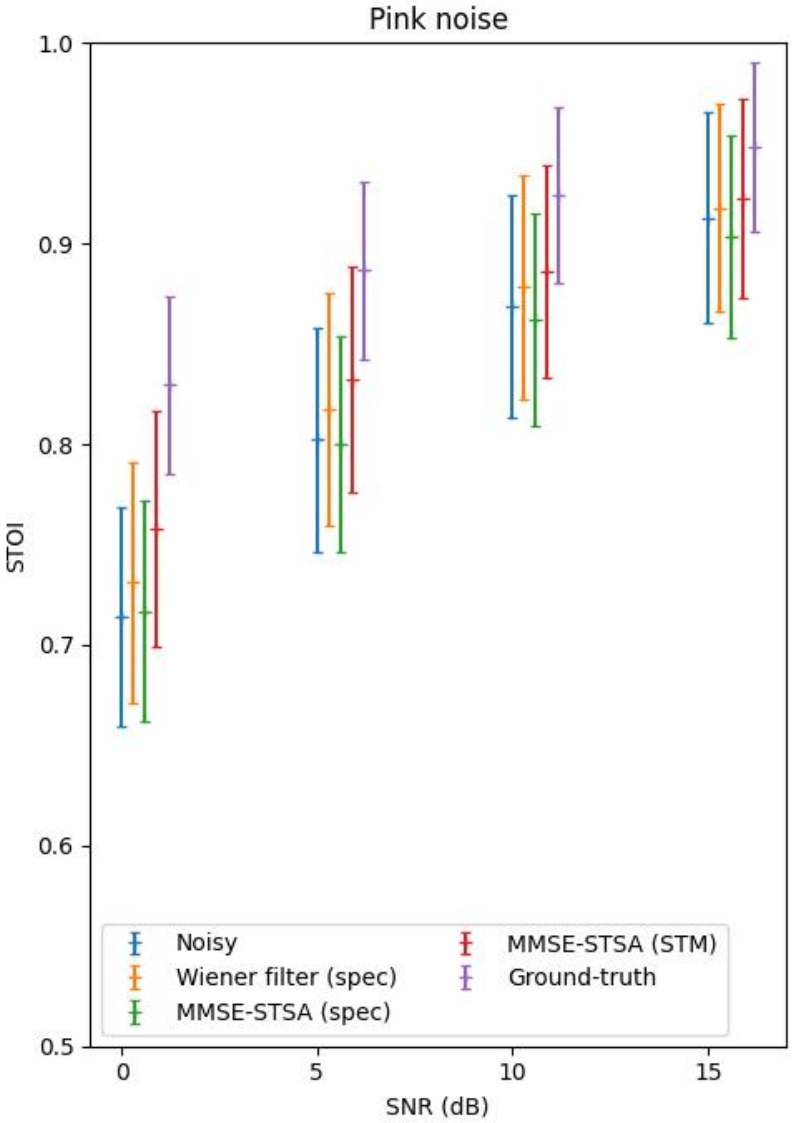


Figure 5.8: STOI score: pink noise

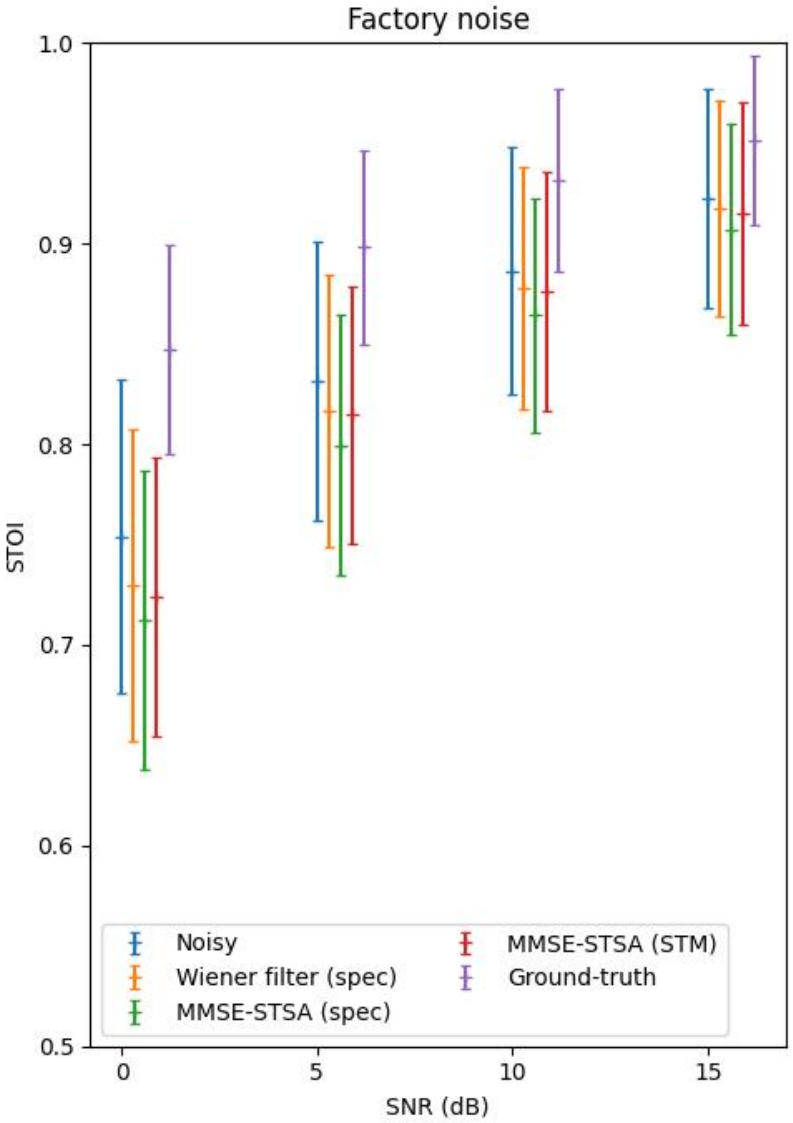


Figure 5.9: STOI score: factory noise

SNR scores, especially under challenging noise conditions.

While the segmental SNR assesses the quality of the enhanced audio based on distortion, the PESQ score evaluates audio quality by simulating the human auditory system. Based on the result in Figs. 5.4 to 5.6, the proposed noise suppression method could enhance the quality of the noise-suppressed audio in comparison with the noisy audio. However, the PESQ score of the MMSE-STSA with STM feature is lower compared to the benchmark methods, indicating the relatively lower quality score of the noise-suppressed audio. It is crucial to consider that the ground-truth score (the one in purple color) is inherently lower as compared to the other algorithms.

The STOI score evaluates the intelligibility of the enhanced audio. The evaluation scores shown in Figs. 5.7 to 5.9 indicate the enhanced intelligibility of the audio obtained using the proposed method, achieving the best score compared to the other algorithms. Nevertheless, it is intriguing to observe that the STOI score for the factory noise, which is a non-stationary noise, remains similar to the one noise-suppressed by the Wiener filter.

The discrepancy in evaluation scores between speech quality and intelligibility metrics indicates that the proposed method provides a more balanced result by improving speech quality while retaining or improving speech intelligibility.

The subsequent evaluation involves a comparison between the spectrograms of the noise-suppressed speech using the MMSE-STSA with STM feature against benchmark methods. The spectrogram comparison of noise enhancement uses noisy speech in white noise at 5 dB. Observations reveal that the enhancement result obtained by the proposed method preserves the vertical structure, resembling the form of white noise, which is also evident in the clean spectrogram. However, the enhanced audio by the proposed method lacks fine structure.

In comparison with the result obtained with the Wiener filter, the result from the proposed method exhibits fewer "small dots" in the higher frequency area, indicating reduced musical noise in the noise suppression result. The noise suppression result achieved with the MMSE-STSA using spectral features also exhibits diminished musical noise, resulting in an overall cleaner output. However, this noise suppression may lead to decreased speech intelligibility, as indicated by the STOI score.

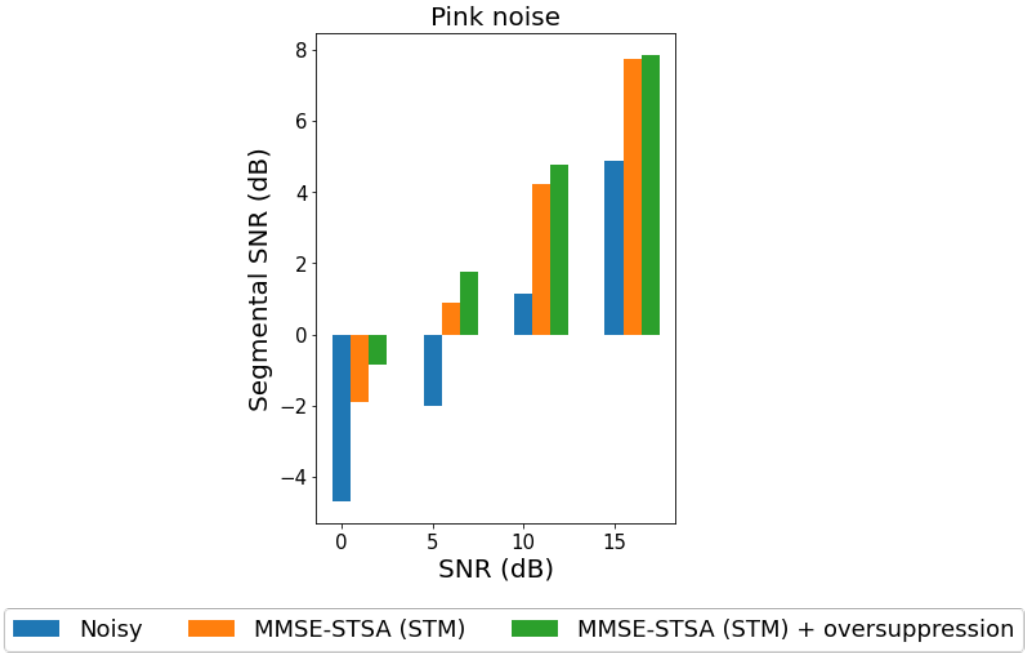


Figure 5.10: Noise Suppression Result with STM

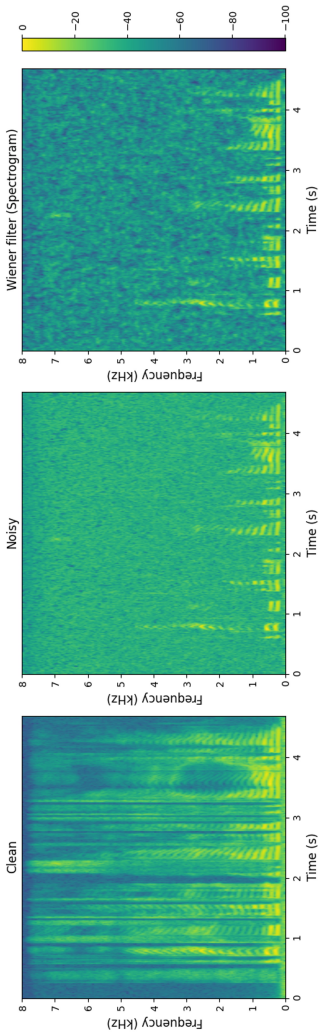


Figure 5.11: Spectrogram of enhanced speech by Wiener filter (spectral feature)

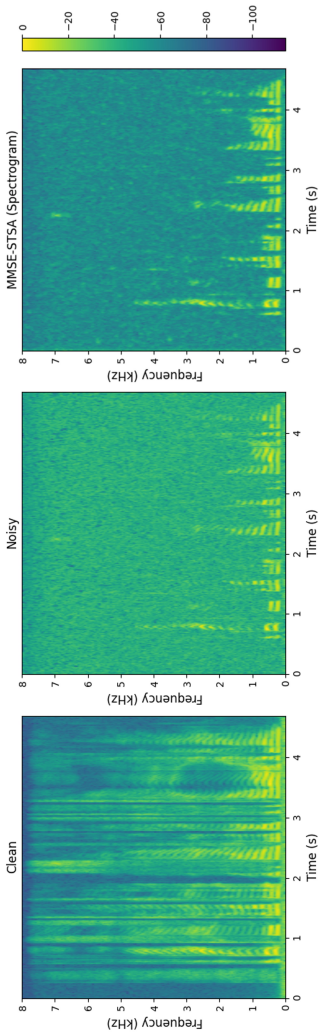


Figure 5.12: Spectrogram of enhanced speech by MMSE-STSA (spectral feature)

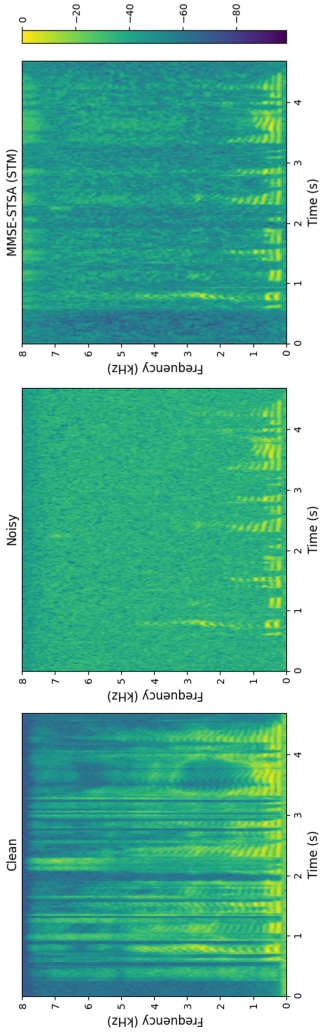


Figure 5.13: Enhanced spectrogram by MMSE-STSA (STM feature)

Chapter 6

Conclusion

6.1 Summary

Noise suppression is a critical area of research in audio processing that aims to remove unwanted noise whilst preserving the desired quality of the noise-suppressed audio. This study involves an investigation into an alternative approach aimed at enhancing noise suppression techniques through the incorporation of alternative features. These feature modifications are designed to integrate adverse side effects and enhance the discernibility between speech and ambient noise, thus yielding more favorable outcomes in the context of noise reduction.

The proposed noise suppression algorithm is the extension of the MMSE-STSA algorithm to the STM domain. This method leverages the knowledge of noise and speech characteristics in the joint STM domain to achieve more controlled and effective noise reduction. During the implementation, this study explores the over-suppression technique to enhance the noise suppression result.

The research evaluates the proposed method using objective metrics, including segmental SNR, PESQ, and STOI. The proposed method shows promise in improving the intelligibility of the enhanced audio and outperforms other algorithms in STOI. However, it does not achieve the best overall performance in PESQ and the segmental SNR compared to spectral domain-based noise suppression methods. After further investigation, it was found that the reason for this result is due to the greater importance of phase information in the STM domain compared to the spectral domain. Therefore, it is worth exploring the integration of the phase information further for a more robust noise suppression using STM.

Acknowledging its limitations, such as its scope for single-channel uncorrelated noise suppression and the substantial role of phase information in the STM domain, this research underscores the potential for future advancements in noise suppression algorithms and audio processing techniques.

6.2 Research contribution

The main contribution of this study focused on a comprehensive exploration of the STM feature as an alternative approach to noise suppression. As one of the early research on the usage of the STM feature for noise suppression, this study contributes to the new findings about the characteristics of speech and noise in the STM domain. By investigating these characteristics, valuable insights into the way STM-based noise suppression can be implemented to reduce background noise effectively.

Based on the knowledge of how noise and speech behave in the joint STM domain, a proposed noise suppression method that incorporates and considers the suppression of noise while maintaining the quality of speech is explored. Therefore, based on the result of this study, it is expected that further advancement can be studied to explore the possibilities of the STM feature in noise suppression or audio processing in general.

6.3 Remaining works

In this section, several remaining works are described.

1. Consideration to include the phase information for noise suppression. As mentioned in the proposed method, the modification of the STM feature is only applied to the amplitude value while keeping the noisy phase information. As seen from the evaluation score of the ground-truth result of the proposed noise suppression, the differences between the suppressed audio and the original clean signal are still considerably high. The noisy phase information causes this big difference in the ground-truth result, and the effect is bigger compared to the implementation of the noise suppression in the spectral domain. Therefore, considering another method to include the phase information for noise suppression is necessary to increase the result using the STM feature presented in this study.
2. Consideration of a finer STM frequencies partition. The current partition of modulation frequencies is derived solely from the observation of noise and speech characteristics. To enhance the noise suppression results, the number of groups and the partition points can be treated as variables that hold potential for optimization.
3. Incorporating other noise suppression method which is more robust for non-statistical noise. The MMSE-STSA algorithm adopted for the STM feature in this study is developed assuming a more stationary

tendency from the noise than that of the speech. This algorithm also adopts the general assumption in which the noise and speech have a distribution close to the Gaussian distribution. Hence, exploring other noise suppression methods which incorporate non-stationary noise, such as data-driven machine learning techniques, might be beneficial to enhance the result of this study.

References

[1] N. Bryan, D. Sun, and E. Cho. (2013) Single-channel source separation tutorial mini-series. [Online]. Available: https://ccrma.stanford.edu/~njb/teaching/sstutorial/

[2] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, p. e1000302, 2009.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[4] R. Bentler and L.-K. Chiou, "Digital noise reduction: An overview," *Trends in Amplification*, vol. 10, no. 2, pp. 67–82, Jun. 2006. [Online]. Available: https://doi.org/10.1177/1084713806289514

[5] S. D. Soli and L. L. Wong, "Assessment of speech intelligibility in noise with the hearing in noise test," *International Journal of Audiology*, vol. 47, no. 6, pp. 356–361, Jan. 2008. [Online]. Available: https://doi.org/10.1080/14992020801895136

[6] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: The neighborhood activation model," *Ear and Hearing*, vol. 19, no. 1, pp. 1–36, Feb. 1998. [Online]. Available: https://doi.org/10.1097/00003446-199802000-00001

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[9] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[10] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[11] C. V. Pavlovic, G. A. Studebaker, and R. L. Sherbecoe, "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *The Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 50–57, 1986.

[12] D. O'Shaughnessy, "Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[15] N. Wiener, "Time series," 1977.

[16] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[17] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr," *arXiv preprint arXiv:2201.06685*, 2022.

[18] H. Dudley, "The carrier nature of speech, bell system tech," 1940.

[19] "Perception of speech as a modulated signal," in *Proceedings of the Tenth International Congress of Phonetic Sciences*. De Gruyter Mouton, Dec. 1984, pp. 29–40. [Online]. Available: https://doi.org/10.1515/9783110884685-006

[20] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech communication*, vol. 52, no. 5, pp. 450–475, 2010.

[21] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, 2012.

[22] B. P. Bogert, "The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking," in *Proc. Symposium Time Series Analysis, 1963*, 1963, pp. 209–243.

[23] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America*, vol. 45, no. 2, pp. 458–465, 1969.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[25] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[26] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, 2000.

[27] M. Ter Keurs, J. M. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception. i," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2872–2880, 1992.

[28] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[29] N. Mesgarani and S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–1105.

[30] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE transactions on Audio, Speech, and Language processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[31] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice-hall Englewood Cliffs, 1999.

[32] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, Feb. 2005. [Online]. Available: https://doi.org/10.1016/j.specom.2004.08.001

[33] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 1979, pp. 208–211.

[34] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[35] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[36] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[37] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, pp. 920–930, 2006.

[38] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 416–422, 2002.

[39] T. Vuong, Y. Xia, and R. M. Stern, "A modulation-domain loss for neural-network-based real-time speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6643–6647.

[40] C.-C. Hsu, T.-H. Lin, and T.-S. Chi, "Fft-based spectro-temporal analysis and synthesis of sounds," in *2011 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5388–5391.

[41] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Spectro-temporal subband wiener filter for speech enhancement," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4001–4004.

[42] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, vol. 8, 2016, pp. 352–356.

[43] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth international conference on spoken language processing*, 1998.

[44] S. Quackenbush, T. Barnwell, and M. Clements, "Objective measures of speech quality. prentice hall," *Englewood Cliffs, NJ*, 1988.

[45] E. Onggosanusi, B. V. Veen, and A. Sayeed, "High throughput wideband space-time signaling using channel state information," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. IEEE. [Online]. Available: https://doi.org/10.1109/icassp.2001.940489

[46] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[47] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[48] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6405–6409.

[49] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech,"

IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[50] S. Ravuri and S. Wegmann, "How neural network features and depth modify statistical properties of HMM acoustic models," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016. [Online]. Available: https://doi.org/10.1109/icassp.2016.7472645