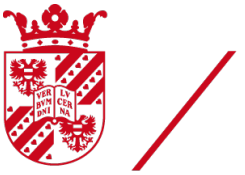
A Self-Supervised Approach to Speech Enhancement in Noisy Climbing Gym Environments



university of groningen

campus fryslân

Thesis

For the fulfillment of the MSc Voice Technology

by

Ellemijn Galjaard

primary supervisor: Dr. Shekhar Nayak
external supervisor: Dr. Nitya Tiwari
second reader: Dr. Matt Coler

Keywords: speech enhancement, speech denoising, single-channel enhancement, environment noise, climbing gym noise

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my external supervisor, Nitya Tiwari, for her guidance throughout this project. Her help has been invaluable, and I could not have completed this project successfully without it. I would also like to express my gratitude to Shekhar Nayak and Vass Verkhodanova, who came up with the research topic and greatly encouraged me to pursue this project. Special thanks to Elja Leijenhorst, who was kind enough to share her data set with me. She was always patient and willing to answer any questions I had about the data and climbing in general, no matter how confused I was. Finally, I would like to thank all of my classmates in Leeuwarden for the fun social activities and making this year really enjoyable for me. I wish you all the best of luck with your future endeavours, and I hope you achieve success in whatever goals you pursue.

CONTENTS

ABSTRACT

Existing speech enhancement models struggle to generalize to diverse acoustic environments with unfamiliar noise types. The acoustic environment of a climbing gym presents a particularly interesting challenge to speech enhancement models due to high levels of complex ambient noise. Therefore, this study investigates the effectiveness of a self-supervised speech enhancement model in removing climbing gym noise from speech signals. In order to achieve this goal, a range of different experiments are conducted which consider various factors that could have an influence on the model's effectiveness, such as variations in training data and the inclusion of the audio signal's phase information during model training. Despite the inconclusive results obtained, this study provides valuable insights into the complexities of speech enhancement tasks. Furthermore, it identifies potential areas for future research that can contribute to developing more effective speech enhancement algorithms for challenging noisy environments.

1

INTRODUCTION

In recent years, speech enhancement has emerged as an important research area within the field of audio signal processing. The main goal of this discipline is to enhance the intelligibility and perceptual quality of speech signals that have been corrupted by various forms of noise. While many speech enhancement models have shown remarkable results in removing noise from speech signals, it is impossible to account for all variations of noise. As a result, these models do not tend to generalize well to unseen acoustic environments, as they struggle to cope with the mixture of unfamiliar noises encountered in such settings.

The acoustic environment of a climbing gym presents a particularly interesting challenge to speech enhancement models. As a facility designed for indoor rock climbing, the climbing gym is often filled with high levels of ambient noise. Moreover, this type of climbing gym noise is highly complex: it is composed of a mixture of both transient and continuous sounds that occur at irregular intervals, such as the clamor of climbing equipment, footsteps, shouting, and drilling noises during climbing route installation. Another defining characteristic of climbing gym noise is its variability. For example, a climber might make a small jump onto a safety mat at the bottom of the wall, which can cause a sudden burst of loud noise. Such sounds are not at all consistent and can vary in their intensity, which can make it difficult for speech enhancement architectures to adapt to the noise characteristics.

Despite the fact that climbing gym environment noise is highly variable and complex, no research has thus far been dedicated to exploring how well speech enhancement models respond to this type of noise. Some popular examples of noise data sets that have been widely used in past speech enhancement studies are AURORA [1], CHiME [2], DEMAND [3], and NOIZEUS [4]. These data sets include noise from suburban trains and train stations, airports, car traffic, restaurants and cafes as well as dinner parties at home, and several diverse indoor and outdoor settings such as offices, kitchens, and parks. When it comes to the indoor settings, the noise tends to be recorded in smaller confined spaces like cafes and meeting rooms. The climbing gym, however, is a far larger and emptier area than these kinds of settings, which influences the way sound travels

around the space. This can lead to sound reflections such as echos or prolonged reverberation, but this also more generally causes the distribution of noise sources (such as conversations or equipment clatter) to be much more spread out across the space. Considering this difference, investigating how enhancement architectures react to this specific type of noise would be a valuable contribution to the field of speech enhancement. This leads us to our research question and hypothesis, stated in the next section.

1.1. RESEARCH QUESTION AND HYPOTHESIS

In order to address the observed gap in the literature, this thesis aims to investigate the effectiveness of speech enhancement architectures in removing climbing gym noise from speech signals. For this purpose, we use a novel data set of real-world climbing gym noise that was recently compiled by Elja Leijenhorst [5]. As the amount of gym noise recordings that was collected is limited, this thesis will employ a self-supervised learning approach to make most efficient use of the data. Self-supervised learning, in this context, simply means that a machine learning model learns to extract useful representations from data without relying on explicit labels or parallel data pairs. It is therefore particularly suitable for cases where data collection is rather costly.

For this reason, this thesis utilizes a self-supervised speech enhancement model. As studies on self-supervised learning for speech enhancement are still relatively limited, this thesis will specifically adopt a model developed by Wang et al. [6], whose research stands out as one of the few studies that have explored a self-supervised approach in this domain. Hence, the main research question which this thesis aims to answer is:

> *How effective is Wang et al.'s self-supervised speech enhancement model in removing climbing gym noise from speech signals?*

To answer this research question, we will first test their model on mixture signals composed of speech and climbing gym noise, and observe how well the model can recover clean (i.e. noise-free) speech signals from these mixtures.

This thesis hypothesizes that Wang et al.'s model may not perform optimally on these mixtures for two reasons. First, the complex and dynamic character of climbing gym noise will likely be challenging to a speech enhancement model unfamiliar with this noise type. Second, Wang et al.'s model only processes the magnitude of the mixture signals (i.e. the strength of the frequency components of the waveform), while completely ignoring the phase information (i.e. information on how the waveform is progressing in its cycle). However, this kind of approach goes against the findings of recent studies [7–10] which have pointed out that ignoring phase information can lead to a limited performance of speech enhancement models. Motivated by this possibility for further improvement, this thesis explores the potential benefits of incorporating information about the phase of the audio signal during the training of Wang et al.'s model, which is an often overlooked component in more traditional speech enhancement algorithms that mainly focus on the magnitude. We hypothesize that doing so will improve the effectiveness of Wang et al.'s model in removing climbing gym noise from speech signals.

Overall, the impact and relevance of this research would lie in its potential to contribute

to a better understanding of speech enhancement in noisy environments and to the development of more effective algorithms for improving speech communication in various contexts, including climbing gyms.

1.2. THESIS STRUCTURE

To establish a foundation for addressing the research question, Chapter 2 will begin by introducing the field of speech enhancement and expanding on relevant literature. After this, Chapter 3 will provide an overview of the data and data pre-processing steps. Chapter 4 discusses the methods used to train and adapt Wang et al.'s model. The results of these experiments will be provided in Chapter 5. Following this, Chapter 6 will summarize the outcomes of this thesis and reflect on these results. Additionally, it will discuss any issues and limitations encountered during this study and propose possible directions for future research.

2

Literature Review

This chapter provides an overview of relevant literature on speech enhancement.[1] *Section 2.1 starts out with a general introduction to the topic. This is followed by a more detailed description of enhancement methods in Section 2.2, which outlines differences between methods that operate in the time-frequency domain versus the time domain. Finally, as this thesis is concerned with self-supervised learning, Section 2.3 is dedicated to comparing supervised, unsupervised, and self-supervised learning for speech enhancement.*

2.1. Introduction to Speech Enhancement

Speech enhancement (SE) is the task of recovering a clean speech signal from a noise-corrupted speech signal, also known as the noisy mixture signal [6, 10, 11]. A common way of explaining it is like an additive sum $x(n) = y(n) + z(n)$, where $x(n)$ is the overall mixture signal with n representing the frame index, $y(n)$ the clean target signal, $z(n)$ the noise or distortion, and the task is to estimate the enhanced $\hat{y}(n)$ from $x(n)$ [7–9, 12, 13]. See Figure 2.1 for a visualization.

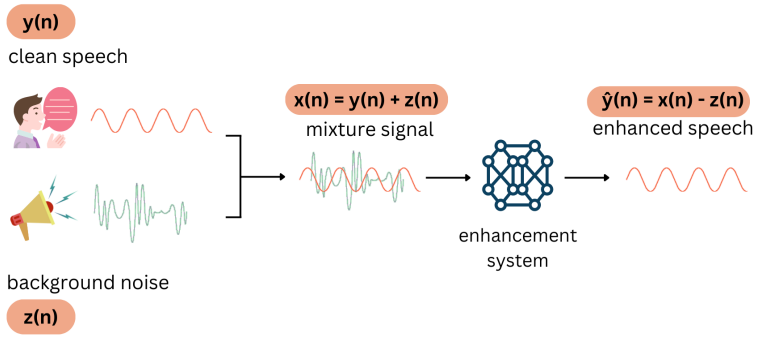


Figure 2.1: Single-channel speech enhancement

[1] Please refer to Appendix A for information on how the resources for this literature review were obtained.

The main motivation for SE is to improve the intelligibility and perceptual quality of speech signals [11, 13–17]. This has multiple practical applications, which include improving the general performance of hearing aids in noisy environments, and optimizing automatic speech recognition (ASR) systems by improving the quality of the audio input [7, 8, 11].

It should be noted that enhancement tasks may include echo removal and dereverberation [14, 18], but this thesis is primarily concerned with background noise suppression, also known as speech denoising. The terms “speech enhancement” and “speech denoising” are therefore used interchangeably throughout this thesis. Generally speaking, there are two main categories of speech denoising methods: so-called *conventional methods* and *deep learning-based methods* [11, 19]. Moreover, there are *hybrid methods* which combine the two aforementioned categories [19]:

1. *Conventional methods* are knowledge-based, meaning that they rely on experts’ a priori “assumptions regarding the statistical characteristics of the signals" [11, p. 1]. Popular conventional methods include techniques like spectral subtraction, Wiener filtering, and Minimum Mean Square Error (MMSE) methods [9, 16, 19]. These methods try to reconstruct the noise and clean speech signals by analyzing their statistical properties. For spectral subtraction, for example, the average of the noise spectrum is estimated at speech pauses, and this is then subtracted from the estimate of the noisy mixture spectrum to obtain the clean target [11, 13]. However, this method makes the assumption that noise is additive and relatively stationary, which is not always the case [13].

2. As different from conventional methods, *deep learning-based methods* do not rely on knowledge-based assumptions. Rather, they are data-driven, meaning that they try to establish nonlinear relationships between the input data – the mixture signals – and the output data – the clean speech signals [11, 19]. These methods utilize neural models such as convolutional and recurrent neural networks to learn a function that maps input to output.

3. *Hybrid methods*, as the name suggests, combine conventional approaches and deep learning. An example of such a hybrid approach can be found in [20], where the authors use a neural network for only those parts of the noise reduction (e.g. complex noise patterns or variations across frequency bands) that are difficult to predict with conventional methods, thereby lowering the number of parameters and complexity of the computations.

This literature review will mainly focus on studies that implement deep learning-based methods, as they are best suited to non-stationary noise [19]. An example of stationary noise is a constant hum or buzz, whereas non-stationary noise changes over time. Climbing gym noise can be considered non-stationary, as typical sounds heard in this environment (such as footsteps or rope movements) tend to be fairly sporadic. Clearly, it is very challenging to develop conventional, knowledge-based methods for this kind of variability within the noise signal – deep learning-based methods, which can learn highly complex nonlinear functions, are better equipped to deal with this kind of sporadicity.

Within the context of speech enhancement or speech denoising, these deep learning-based methods can be further categorized in terms of their operation domain. While sources may slightly differ in exactly how they categorize these methods, this thesis distinguishes between those enhancement methods that operate in the *time-frequency* (TF) domain and those that operate solely in the *time* domain, a categorization that is also followed in [10, 16, 19, 21, 22]. The next section will cover these two categories in more detail.

2.2. SPEECH ENHANCEMENT METHODS: TIME-FREQUENCY VS TIME DOMAIN

Deep learning-based speech enhancement methods can be broadly categorized into two operation domains: the time-frequency (TF) domain and time domain. The time-domain methods take in the waveform input directly, which is what we refer to as the *raw* waveform. TF methods, on the other hand, transform the raw waveform into a TF-based representation for processing, before eventually re-synthesizing this representation back into a waveform. Both these approaches will be discussed in the next subsections.

2.2.1. TIME-FREQUENCY DOMAIN

TF-based enhancement methods typically apply a Short-Time Fourier Transform (STFT) to the raw waveform: the time-domain signal is first divided into overlapping segments (i.e. frames) via a window function, and a Discrete Fourier Transform (DFT) is applied to each frame in order to obtain its constituent frequencies [17]. These frequencies make up a complex-valued spectrum, where the frequency components of the spectrum carry both magnitude and phase information [7, p. 1]. In simple terms, *magnitude* can be seen as the strength or size of the frequency components at specific time points, whereas *phase* can be seen as the relative positioning of those frequency components on the raw waveform – i.e. how the waveform is progressing in its cycle. The phase is therefore often represented as an angle within this cycle, providing a standardized reference for understanding and comparing phase relationships between frequency components. In Figure 2.2, we can see how the phase angle encodes the relationship between the frequency components at two different points on the waveform.

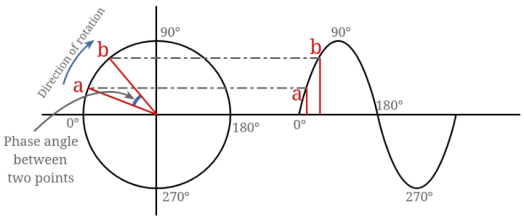


Figure 2.2: Phase angle of points on a sine wave [23]

As the phase is more difficult to process due to its cyclical nature, the phase infor-

mation is often excluded during training [7, 24]. The final TF representation used for training is usually a magnitude spectrogram, which can be seen as a 2D matrix that is obtained by squaring the magnitude information of the STFT. Another example of a TF representation, although less commonly used in speech enhancement, is the mel-frequency spectrogram [25]. Here, the linear frequency scale of the STFT is mapped onto the logarithmic mel scale, which more closely corresponds to the way humans perceive frequency.

The idea behind using such TF-based representations, as opposed to raw waveforms, is that they provide a more detailed overview of auditory patterns, such as "proximity in frequency and time, harmonicity and common amplitude and frequency modulation" [10, p. 3816], which can make it easier for the deep-learning model to learn meaningful features that help filter out the noise components. In other words, the TF representation makes the speech and noise patterns in the signal more easily distinguishable [22]. For a visual aid, see Figure 2.3, which clearly showcases the richness of information contained in a TF representation as compared to the original time-domain waveform.

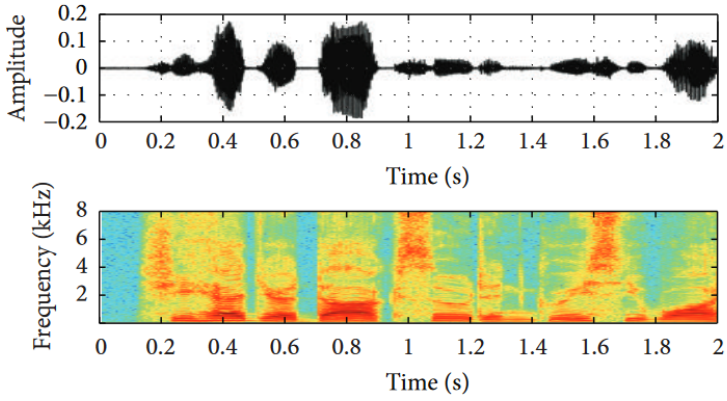


Figure 2.3: Waveform (top) and magnitude spectrogram (bottom) of the same audio signal [26, p. 7]

So how do these TF models work? As discussed previously, deep-learning models that operate in the TF domain first take the noisy waveform as input, and transform this input into a TF-based representation such as the noisy magnitude spectrogram [16, 22]. They then start learning the underlying patterns of this input to be able to identify and separate the noise components from the desired speech components, thereby estimating the clean TF representation. This clean representation is then synthesized into a clean waveform using an inverse Short-Time Fourier Transform (STFT), which merges and transforms the frames back into a time-domain representation [7, 16]. For a visual depiction of a standard TF-domain enhancement framework, please refer to Figure 2.4.

It is important to note that the final clean waveform cannot be synthesized based on the magnitude spectrogram input alone, as waveform reconstruction also requires information about the phase of the waveform. However, as mentioned previously, phase estimation can be quite difficult — instead of estimating the phase, a common workaround is therefore to simply combine the estimated clean magnitude with the phase informa-

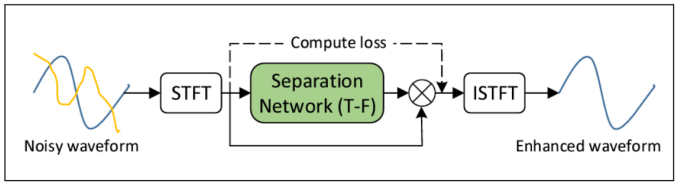


Figure 2.4: A typical TF-based speech enhancement framework [10]

tion from the noisy input [7].

Indeed, many TF-based speech enhancement approaches choose to re-use the phase from the noisy input rather than predicting it. While older research [27] claimed that phase estimation does not improve the predicted signal, more recent research [7–10] has shown that it can be beneficial to also estimate the phase information of the speech signal, and simply re-using the noisy phase from the input can in fact "create a performance upper bound" [10, p. 3816]. Predicting the speech signal's phase can particularly contribute to enhancing the *intelligibility* of the signal (Ibid.). However, phase estimation in the TF domain is a rather difficult task, due to the fact that the phase spectrogram (which is obtained by computing the angle of each element in the STFT matrix) is highly unstructured as compared to its magnitude counterpart [24, 28], as seen in Figure 2.5. Although the phase is inherently cyclical and its values are continuous, the phase angles tend to be wrapped within the range of $-180°$ and $180°$ so that they are easier to process, and this wrapping is often seen as the reason for the unstructured character of the phase spectrogram [24].

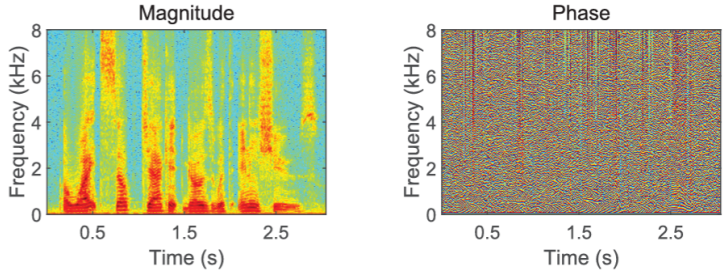


Figure 2.5: Examples of magnitude and phase spectrograms [24, p. 484]

Because phase estimation is difficult, there is still a large amount of projects which focus on magnitude estimation only [7]. We can therefore distinguish between so-called *TF-magnitude* and *TF-complex* approaches, where *TF-complex* approaches leverage magnitude as well as phase information [21].

There are multiple ways in which a TF-complex approach can incorporate phase information. One approach involves predicting the magnitude and phase information of the clean target separately, as is done in Microsoft's model PHASEN [22]. To deal with the lack of structure in the phase spectrogram, PHASEN incorporates information exchange

between the two different streams, meaning that the model can leverage the structural information of the magnitude spectrogram when predicting the phase (Ibid.).

Another way a TF-complex model can incorporate phase information is by directly estimating a complex-valued spectrogram where magnitude and phase are no longer decoupled, thereby avoiding having to estimate the phase separately. To better understand this technique, it is first necessary to briefly explain how we can represent the Short-Time Fourier Transform (STFT) of a noisy mixture signal using different coordinate systems. When representing the STFT of each time-frequency bin in the noisy input spectrogram, past studies have commonly employed polar coordinates (i.e. the magnitude and phase angle) [24]. As an alternative, more recent work has investigated how we can express the STFT of each TF unit by using a Cartesian coordinate system with a complex exponential (Ibid.). Please refer to Figure 2.6 for a side-by-side comparison of these two coordinate systems.

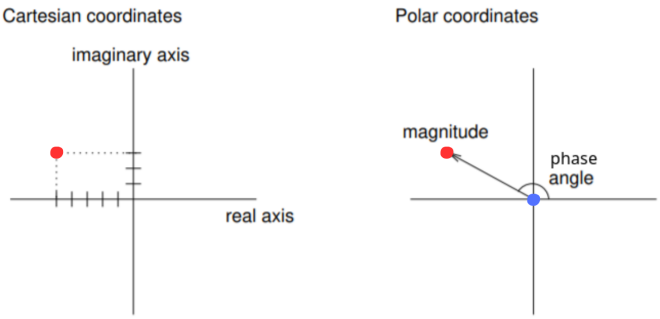


Figure 2.6: Cartesian coordinate system with complex exponential (left) vs polar coordinate system (right) [29]

On the left side, we have the Cartesian coordinate system with the complex exponential, where the x-axis represents the real part and the y-axis represents the imaginary part of the complex number. In this system, a complex-valued TF unit is expressed by its coordinates on these axes. On the right side, we have the polar coordinate system, where a complex-valued TF unit is represented through means of a reference point (the origin at 0,0), where the *distance* of this TF unit from the reference point indicates the magnitude component, and the *angle* it makes with the positive x-axis represents the phase component of this TF unit.

In the Cartesian representation, each TF unit is represented by its real and imaginary

components.[2] By adopting the Cartesian coordinate system with the complex exponential instead of the commonly used polar coordinate system, we can represent the TF units in a more structured and clearer manner. The benefit of this, according to [24], is that we no longer have to deal with the unstructured character of the phase spectrogram, meaning that the phase information does not need to be discarded. Instead, as the TF unit is now represented by its real and imaginary components, we can compute a real and imaginary spectrogram which both exhibit a clear structure, as seen in Figure 2.7.

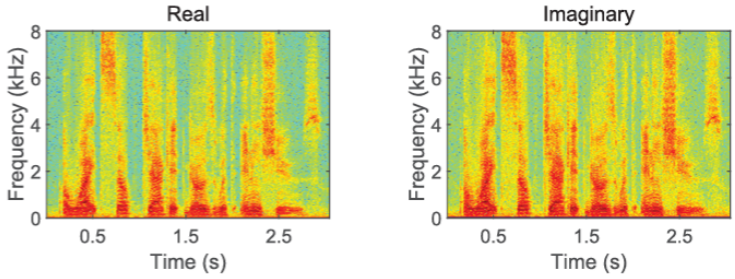


Figure 2.7: Examples of real and imaginary spectrograms [24, p. 484]

Examples of models that estimate the complex-valued spectrogram in this way are the well-known Deep Complex U-Net (DCU-Net) [7] architecture and some of its more recent adaptations like DCRNN [30]. However, it should be noted that DCU-Net is a *masking* model rather than a *mapping* model like most of the models that have been discussed thus far. While *mapping* models directly map noisy input to clean targets, for *masking*-based architectures, the training target is not the clean spectrogram itself but rather an "intermediate mask" [19] that indicates the presence or absence of noise at each TF unit of the spectrogram (i.e. a weighted matrix). Applying this intermediate mask to a noisy spectrogram representation is then supposed to yield a clean estimate [21]. TF-magnitude approaches commonly apply the Ideal Binary Mask (IBM), which assigns 0 or 1 to each TF unit to indicate the presence or absence of noise, or the Ideal Ratio Mask (IRM), which works with probability scores instead of binary values and has proven to be more effective [19, 21]. TF-complex approaches, like DCU-Net and DCRNN, commonly apply the so-called complex Ideal Ratio Mask (cIRM) instead, which estimates the ideal speech-to-noise ratio at each TF unit along a complex Cartesian coordinate system [21].

It is important to note, however, that the discussed TF-complex approaches are not the only way of addressing the phase estimation problem. Time-domain approaches also offer a viable option. In fact, the main advantage of using time-domain approaches over TF approaches is that the phase and magnitude are not separated to begin with, as

[2] This thesis will not go too far into the mathematical theory behind this concept. The most important thing to understand is that these real and imaginary parts retain both magnitude and phase information by using a different spatial representation. Specifically, the magnitude can be determined by taking the square root of the sum of the squares of both components, i.e. $(\texttt{sqrt(real}^2\texttt{+imaginary}^2\texttt{)})$. The phase can be determined by taking the inverse tangent (arctan) of the imaginary part divided by the real part, i.e. `arctan(imaginary/real)`, which produces an angle representing phase information [24].

will be discussed in the next subsection.

2.2.2. TIME DOMAIN

As mentioned previously, time-domain approaches to speech enhancement work with raw waveform input. Similar to TF-based approaches, they first divide the signal into frames using a window function. However, unlike TF-based approaches, time-domain methods do not apply any kind of Fourier transform to these frames. This means that they do not separate the magnitude and phase components of the signal, nor do they require complex phase estimation. A great advantage of this is that time-domain approaches are typically less computationally expensive [17].

The big downside to time-domain methods, on the other hand, is that they generally perform better with larger frame sizes [31, 32], which leads to a higher number of parameters and a larger model size [17]. While TF-based approaches are confined to capturing only those features that are explicitly present in the TF representations, time-domain approaches have access to the original waveform from which they can derive features autonomously [17]. This means that, if the window size is too small, the input might not contain enough important information for the time-domain architectures to learn meaningful features [31].

Despite these disadvantages, there are examples of time-domain architectures that achieve high performance. According to [10, 22], two of the most influential time-domain models that have been developed in recent years are SEGAN and Conv-TasNet. SEGAN [33] is an end-to-end model that was one of the first to use generative adversarial networks (GAN) for speech enhancement. This GAN is composed of a generator, which produces an enhanced or denoised version of noisy input speech, and a discriminator, which determines whether this enhanced version can be classified as clean speech or not. In this way, the generator and discriminator are able to mutually optimize each other.

Conv-TasNet [34] was originally designed for speech separation (e.g. to separate speakers in multi-speaker scenarios) but later adapted to speech enhancement [22]. This network is composed of an encoder, a separation network, and a decoder. The encoder applies 1D convolutions to the raw waveform to obtain a feature representation of the entire signal. The separation network uses this representation to estimate a multiplicative function (mask) for each individual speech signal at each time step, and these individual speech signals are reconstructed by the decoder. According to [34], Conv-TasNet outperforms several earlier speech separation models that operate in the TF domain.

Although Conv-TasNet works well for speech separation, the authors in [22] state that when this model is implemented for speech enhancement, "the 2ms frame length appears to be too short" for achieving a similar performance (p. 3). While there are adaptations of Conv-TasNet which make use of longer frame lengths (like TCNN [35]), [22] claims that time-domain methods are simply not as well suited to the speech enhancement problem as TF methods. While operating in the time domain avoids the issues associated with phase estimation, it still seems to be more beneficial to transform the audio signal to a TF representation, as this representation makes the speech and noise patterns within the signal more easily distinguishable to the model [22].

Having discussed some of the differences between speech enhancement methods in

the TF domain and time domain, it is essential to now consider the learning paradigms employed to tackle the enhancement task effectively. Given that this thesis utilizes a self-supervised model developed by Wang et al. [6], as mentioned in the introduction and further explained in Chapter 4, the following section is dedicated to comparing supervised, unsupervised and self-supervised learning for speech enhancement. Moreover, it explains the rationale behind adopting a self-supervised framework.

2.3. COMPARING SUPERVISED, UNSUPERVISED AND SELF-SUPERVISED LEARNING FOR SPEECH ENHANCEMENT

Supervised learning is by far the most common in speech enhancement [11]. In supervised approaches, the deep-learning models are trained to estimate the clean targets by learning from parallel pairs of noise and clean speech data. *Parallel*, in this case, means that the noise and speech have been recorded in the same acoustic environment. These models learn by comparing the predicted clean speech signal (also known as the *enhanced* signal) with the actual clean speech signal from the data pair. By computing the loss based on this difference, the model can be optimized for better performance. Some of the more popular models discussed in this chapter, like Deep Complex U-Net [7], Conv-TasNet [34] and SEGAN [33], are trained in a supervised manner. While these models have achieved good results, there are some clear disadvantages to supervised learning, as discussed in [6, 36]:

1. Obtaining the parallel noisy-clean data pairs for supervised learning can be rather expensive. Moreover, it is difficult to ensure that the collected clean targets are truly 'clean' and not at all contaminated by noise.

2. Supervised models may not respond well to noise in a real-world, uncontrolled environment because it is variable and unpredictable. Even if the model is deployed in a similar environment to the one it was trained in, the model might still encounter unseen noises to which it cannot adapt itself very effectively.

3. Related to the previous point, supervised models have limited generalizability to new or different acoustic environments. If the model is deployed in a different environment, a significant drop in performance is likely to occur.

Seeing as there are disadvantages to supervised learning in terms of data collection and generalizibility, in recent years some research has been dedicated to unsupervised and self-supervised learning for speech enhancement.

Unsupervised learning relaxes the constraints on the data collection by training models using either non-parallel pairs of clean and noisy data, clean data alone, or noisy data alone [37]. Unsupervised *noise-dependent* methods learn the noise characteristics of the noisy samples during training, while *noise-agnostic* methods solely rely on clean speech signals during training and estimate the noise characteristics at testing time [36] [37]. A recent example of an unsupervised model which utilizes a noise-dependent or "noise2noise" method is found in [38]. Rather than mapping the noisy input samples to their corresponding clean target samples, as is done in supervised learning, the authors train their model to map noisy input samples to uncorrelated noisy target samples

(Ibid.). As the input and output have different noise characteristics, the model effectively learns to "denoise" the input samples. This lack of correlation between input and output equally ensures that the model learns to denoise data samples in a more generalized manner, rather than learning a mapping from one specific noise type to another (Ibid).

An example of an unsupervised noise-agnostic method can be found in [39]. The researchers first train a variational autoencoder (VAE) on only clean speech signals in order to learn their underlying patterns and characteristics. At testing time, the VAE uses this knowledge to approximate clean speech from the noisy input. During this step, the VAE is assisted by a more traditional parametric noise model, which estimates the characteristics of the noise present in the input. The parameters of this model are computed using an expectation-maximization algorithm, but the iterative nature of this algorithm (i.e. it continuously updates the parameters) makes such an unsupervised approach rather slow and computationally expensive during the testing stage (Ibid.). Although the authors improve on this issue by developing a more efficient data sampling method, this kind of unsupervised approach is likely not ideal for this thesis.

An alternative learning paradigm is self-supervised learning. Unsupervised and self-supervised learning are similar in that they do not require parallel pairs of noisy and clean data. However, there is a difference in how each approach learns from the available data. Unsupervised learning seeks to uncover the patterns of the noise and/or speech data, while self-supervised learning makes use of this observed structure within the data to generate its own training targets – essentially learning from one part of the input to predict another part of the input.

Unfortunately, studies on self-supervised learning for speech enhancement are still relatively scarce [11, 15]. One example of a self-supervised approach can be found in [40]: in this research, the authors make use of a two-step speech enhancement approach to improve a speech recognition system for Arabic. Since Arabic is an under-resourced language from a speech technology perspective, it can be difficult to obtain clean speech samples, which is why only noisy data samples are used for this experiment. As a first step, the authors train an auto-encoder on pairs of noisy speech in an unsupervised way, which helps it learn patterns and features that it is able to utilize for denoising. As a second step, another auto-encoder uses this denoised output as its training targets to learn how to further enhance and remove any remaining noise from the noisy input speech. By employing such a self-supervised approach, the overall model is able to refine its ability to generate clean speech samples.

The approach described in [40] shows why self-supervised learning is an attractive option for this thesis. By training the model on noisy speech and continuously refining its denoising capabilities, the overall model becomes better at handling different types of noise across diverse acoustic environments. As climbing gym noise is highly variable and the available data for this thesis is limited, it is sensible to adopt a self-supervised approach rather than opting for supervised learning. Although the code for this particular project is not publicly accessible, research by Wang et al. [6] uses a similar self-supervised approach to speech enhancement with open-source code.[3]

This thesis will therefore adopt the self-supervised model developed by Wang et al. [6]. However, one downside to this model is that it only operates on the noisy magni-

[3]The code for this model is available here.

tude spectrogram, and re-uses the phase information from the input noisy speech. As discussed in Section 2.2, phase estimation can improve the intelligibility of the predicted clean speech. Therefore, next to testing the performance of this model on climbing gym noise, this thesis aims to see whether incorporating phase information could help improve the results. The exact architecture of this model and the intended modifications will be described in the methodology, Chapter 4. Before this, Chapter 3 will discuss the data sets used for training the model.

3

DATA

This chapter provides an overview of the data used for this thesis. It is divided into three parts. Section 3.1 will give a description of each of the individual data sets. Section 3.2 will discuss the pre-processing steps that were taken to make the data sets suitable for this research. Finally, Section 3.3 will briefly expand on the ethical considerations surrounding the collection and usage of the data.

3.1. DATA DESCRIPTION

This thesis utilizes three different data sets: one data set containing clean speech audio, and two different data sets containing noise. The clean data set that is used is the `train-clean-100` subset from LibriSpeech [41]. LibriSpeech is an open-source English speech data set and its clean subsets are widely used for speech enhancement tasks. As for the noise, two different data sets are employed: an open-source data set called UrbanSound8K [42] and, most importantly, a newly compiled data set of climbing gym noise recordings. These recordings were made by Elja Leijenhorst [5], who was kind enough to share the data with us for this research.

While this thesis is mostly concerned with how well Wang et al.'s speech enhancement model is able to remove the collected climbing gym noise from noise-corrupted speech signals, the UrbanSound8K data is necessary as we first want to train a baseline version of their model with similar types of noises from a different source. This baseline model will be trained with mixtures composed of clean LibriSpeech samples and UrbanSound8K noise, and tested on mixtures composed of clean LibriSpeech samples and climbing gym noise. The reason for establishing such a baseline model is that it allows us to evaluate how effective the model already is at removing climbing gym noise from speech signals, even when it has not been explicitly trained on this specific noise type. It also allows us to evaluate the performance of any additional experiments through a comparative analysis. A detailed description of the baseline and other experiments will be provided in the next chapter, Chapter 4. The next sections are dedicated to describing the three data sets in more detail.

3.1.1. LIBRISPEECH

LibriSpeech is a widely known audio data set consisting of "read speech" data, meaning speech that is read aloud from text instead of being derived from natural conversation. This set was compiled by having speakers read out excerpts from LibriVox audio books [41]. While the overall data set comprises ca. 1000 hours, LibriSpeech has a few smaller subsets of "clean" speech data. The recordings in these subsets have a cleaner audio (i.e. less background noise) and all speakers within these sets have a similar US English accent (Ibid.). The specific subset which is used for this thesis project is called `train-clean-100`, which contains 100 hours of clean audio data from 251 different speakers. Among these speakers, 125 are female and 126 are male, with ca. 25 minutes of audio data for each speaker (Ibid.). Each file in this set is a .flac type and has a 16000 Hz sampling rate. The duration of each file varies, but the files tend to be less than 30 seconds long.

3.1.2. CLIMBING GYM NOISE

The data set containing the gym environment noise was compiled by Elja Leijenhorst [5] and is currently not publicly available. The data recording took place at a climbing gym in Leeuwarden called *Klimcentrum Noardwand* [43] between January and April of 2023, at various times and various locations within the gym. Each of the files varies in length, ranging from ca. 6 minutes to as long as ca. 2,5 hours, and has a sampling rate of 44100 Hz.

Next to providing information about the files and recording process, this section will also very briefly discuss the acoustic diversity within the climbing gym, as the aim with the recordings was to collect as much and as many types of environment noise within the same gym environment as possible. Figure 3.3 provides a floor map that illustrates the different microphone locations used for recording. As this map shows, there are several separate areas or zones to be distinguished within the gym in terms of their acoustics, such as the lead climbing area, the clip 'n climb, and the top roping area. Figure 3.4 includes pictures of the different areas.

Although situated in the same climbing hall, each of these areas exhibits distinct acoustic characteristics. For instance, the clip 'n climb area is meant for short climbing challenges that are accessible to children – sounds heard in this area can therefore include children shouting or laughing, the clipping and unclipping of carabiners, or children walking on the rubber mats at the bottom of the climbing walls. In the top roping area, common sounds can include the sliding of ropes through belay devices, and the shouting of commands between the climber and the belayer – i.e. the person in charge of the safety rope (see Figure 3.1).

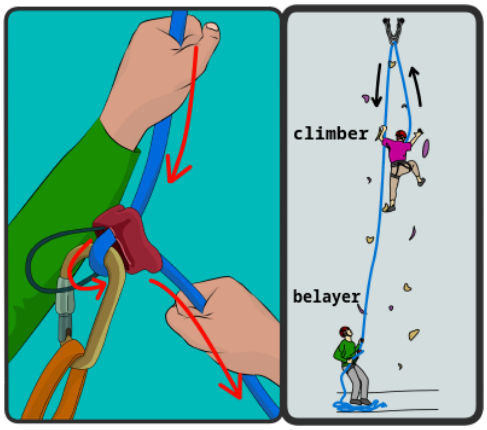


Figure 3.1: Belay device (left) and belaying demonstration (right) [44]

In the lead climbing gym area, the climber is also assisted by a belayer who ensures the climber is safely secured. However, the difference between lead climbing and top roping routes is that the rope is not attached to the top of the climbing route, but the climbers themselves are responsible for pushing the rope through "quickdraws", which are kind of like metal carabiners attached to the wall. As mentioned by [5], these quickdraws make a "sharp clipping sound" which is not heard in the top roping area. See Figure 3.2 for the difference between top roping and lead climbing.

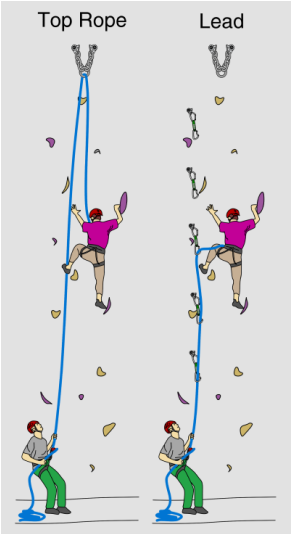


Figure 3.2: Top roping (left) versus lead climbing (right) [44]

Although all recordings were made within the same gym environment, it has been demonstrated that specific areas within this environment can feature certain sounds more or

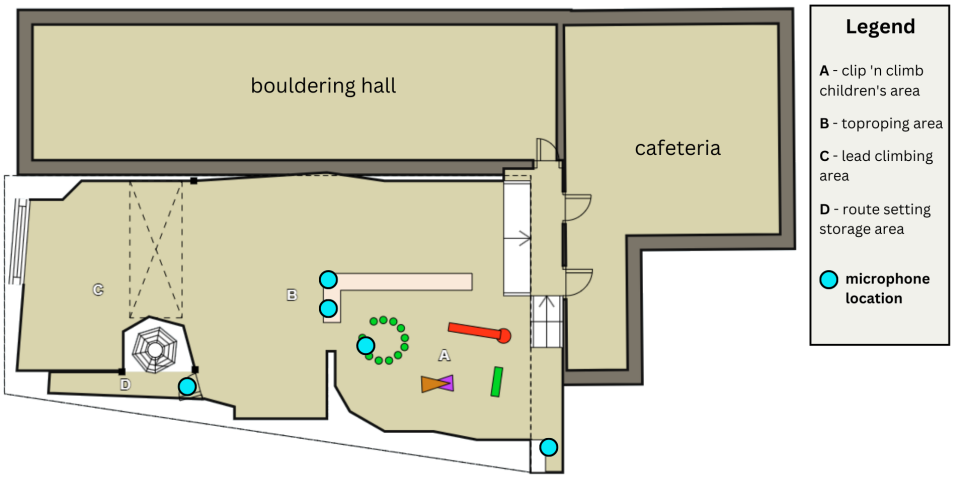


Figure 3.3: Floor map of the climbing gym, created by Elja Leijenhorst [5]

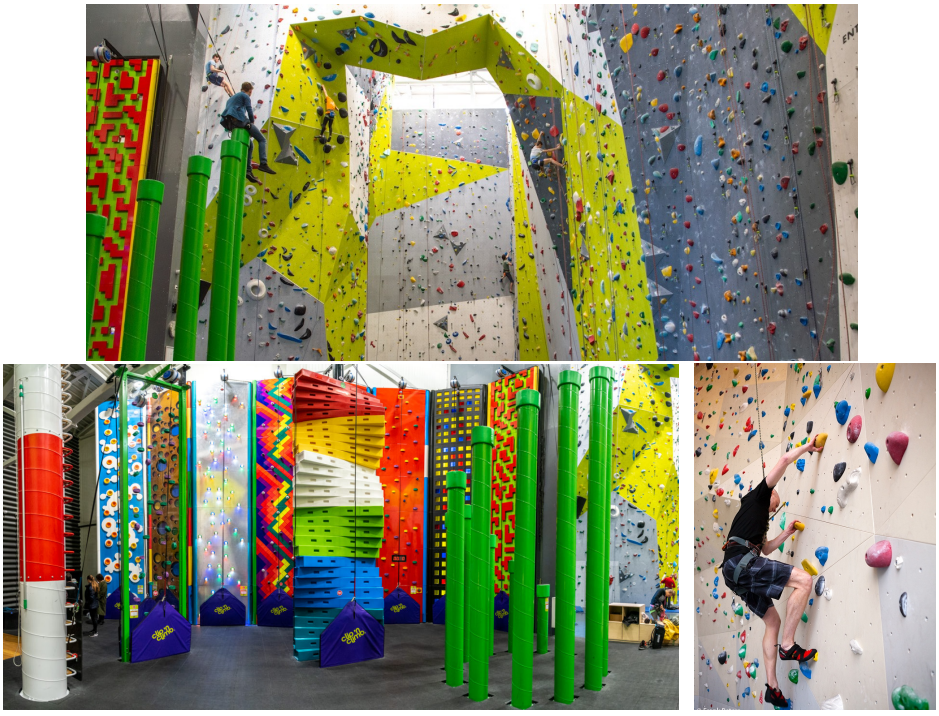


Figure 3.4: In consecutive order: lead climbing area, clip 'n climb, top roping area [43]

less prominently than others. This is why the recordings were made in various different areas of the climbing gym.

However, there is also a separate bouldering gym (see Figure 3.3) which was not considered for this data set, as there is a risk that these sounds are *too* distinct from the environment noise within the main climbing gym. For example, as bouldering is a type of free climbing, there will be a lack of climbing equipment sounds that are present in all of the other recordings. The shapes of the two halls are also different. While the sounds of the bouldering hall were not included in the data for this thesis, future projects could potentially research how the acoustics of the bouldering gym compare to the climbing gym.

3.1.3. URBANSOUND8K

The third data set utilized in this thesis is UrbanSound8K. UrbanSound8K is an open-source data set consisting of 8732 audio excerpts with a duration of max. 4 seconds each [42]. Each of these files contains urban noise from one of the ten different urban noise classes in the data set, which are described in Table 3.1. As the audio in these files is derived from the FreeSound project [45] (which is an online collaborative project where any user is free to upload audio snippets),the sampling rate per file can vary.

class	urban noise type
0	air conditioner
1	car horn
2	children playing
3	dog barking
4	drilling
5	engine idling
6	gun shot
7	jackhammer
8	siren
9	street music

Table 3.1: UrbanSound8K: 10 different urban noise classes [42]

As mentioned previously, the UrbanSound8K data will be utilized to train a baseline model. This model will be trained with mixtures of clean speech and UrbanSound8K noise, and tested on mixtures of clean speech and climbing gym noise. We have therefore selected data from only those urban noise classes that share some resemblance to sounds commonly heard in a climbing gym. The selected noise classes for the baseline model are class **0**, **2**, **4** and **9**.

- **Class 0** was chosen as there are air conditioners in the climbing gym, and this sound is likely to be heard in the recordings.
- As mentioned in the data description, part of the audio was recorded in the clip 'n climb area of the climbing gym. This is an area with shorter climbing challenges that are meant for children. **Class 2** was selected as the sounds of children playing is very similar to some of the noise heard in these recordings.

- **Class 4** consists of drilling noises, which might at first glance not be very relevant to a climbing gym area. However, in a climbing gym, drilling noises are actually heard very frequently as a result of *route setting*. Route setting means that the handholds and footholds on the climbing gym walls are constantly moved around in order to create new routes for the climbers, as can be seen in Figure 3.5.
- **Class 9** was selected as there is often music playing within the climbing gym. It is important to note, however, that street music will likely have slightly different acoustics than music that is played in an indoor gym.

Figure 3.5: Route setting in a climbing gym [46]

Now that we have discussed the data sets selected for this project, the next section will be dedicated to explaining the data pre-processing steps.

3.2. DATA PRE-PROCESSING

Different pre-processing steps were carried out for each of the three data sets.

⇨ **LibriSpeech.** For the LibriSpeech data, the .flac files were first converted into .wav files. As this research does not require all 100 hours of data spoken by 251 speakers, since this would take up more memory and computational power, a smaller selection was made. In the end, 15 female speakers and 15 male speakers were randomly selected, which totals to 30 speakers with around 12.5 hours of clean speech data. Data from 24 of those speakers (12 female and 12 male) were reserved for training, and data from 6 speakers (3 female and 3 male) were reserved for testing purposes.

⇨ **Climbing gym noise.** The recorded climbing gym noise data has a sampling rate of 44100 Hz. This data, in accordance with the clean LibriSpeech data, was first resampled to 16000 Hz via FFmpeg, an audio processing tool that can be run from the Linux commandline. As the recorded noise files were rather large and unequal in size, FFmpeg was also used to divide each file into ca. 1.5 minute-long clips for more efficient processing. Before doing so, 5 seconds was trimmed from the beginning and end of each file to ensure that the audio did not include any instances of the microphone being activated or deactivated. Finally, Python's split-folders library was used to randomly divide these clips into a training and test set, using

an 80/20% split with a random seed of 1337 (the random seed is added for reproducibility – it ensures the same random selection is made every time).

⇨ **UrbanSound8K.** As explained previously, for UrbanSound8K only the data from class 0, 2, 4 and 9 was utilized, as these noise types share some characteristics with the climbing gym noise. The files from these classes were converted into .wav files and resampled to 16000 Hz in accordance with the other data sets. As before, Python's split-folders library was used to randomly divide the files into a training and test set with an 80/20% split and a random seed of 1337.

3.3. ETHICAL CONSIDERATIONS

Next to the description of the data and data pre-processing steps, is important to consider any potential ethical issues with the data collection and usage. While widely known open-source resources like LibriSpeech and UrbanSound8K may not pose significant ethical concerns, it is important to consider any potential ethical risks associated with the collection and usage of the climbing gym noise data. These recordings are meant to capture ambient noise only, but will very occasionally capture small bits of discernible speech when the climbers are in too close proximity to the microphone. In the large majority of recordings, no discernible conversations can be detected, as it is mostly shouting and background noises. However, there are occasional instances where members of the climbing gym can be heard speaking a few words or small phrases clearly.

As voice recordings are considered biometric data and can potentially be used to identify individuals, it is important to handle these cases with utmost care. In light of this, several measures have been taken to ensure the responsible management and protection of the recorded data. First, the members of the climbing gym were informed of the recordings in multiple ways. During the recording process, notice boards were placed next to the microphone as a way of informing individuals about the purpose of the recordings, when the recordings would be taking place, as well as explanation on how to object to being recorded [5]. Next to that, notifications about the recordings were also circulated in the WhatsApp group of the climbing gym, ensuring that members were informed about the ongoing recording activities (Ibid.).

On top of that, Leijenhorst [5] eventually decided not to make the climbing gym data set open-source, in order to ensure responsible protection of data. While any potential speech could be filtered out of the noise recordings using a Voice Activity Detection (VAD) algorithm, some experimentation showed that these algorithms might not always work perfectly on the noisy data. As the data set is too large to manually anonymize within the limited time span, it was decided to keep the data private for now.

Now that the individual data sets and the pre-processing steps have been discussed and any ethical concerns have been addressed, the next chapter will expand on how exactly the data will be utilized for model training. Specifically, the next chapter discusses the methodology of this research: it will first elaborate on the architecture of Wang et al's self-supervised model, and then go into the experiments which train different versions of this model using the previously discussed data.

4

METHODOLOGY

This chapter provides an overview of the methodology and is divided into three main parts. Section 4.1 will first provide background to the methodology: it explains the architecture and experimental set-up of Wang et al.'s model, which serves as the foundation for this research. Section 4.2 will then expand on three different experiments that are conducted using this model. This section also explains the reasoning behind these model experiments and highlights how each one contributes to answering the main research question of this thesis. Finally, Section 4.3 will discuss the evaluation metrics used to assess the performance of each model.

4.1. METHODOLOGICAL BACKGROUND

Before discussing any of the model experiments, this section will first expand on the architecture of Wang et al.'s [6] model and their experimental set-up for model training. Unlike traditional supervised learning approaches, which rely on *parallel* pairs of clean speech and noise data, this model employs a self-supervised method that relaxes the constraints on the data collection by working with *uncorrelated* speech-noise pairs. This simply means that the speech and noise data have been recorded in different acoustic environments. To make most efficient use of the available data, Wang et al. divide their architecture into two parts: they train one autoencoder for clean speech signals, and one for noisy mixture signals.

1. As a first step, they train a clean autoencoder (CAE) to learn useful representations from the clean speech data in an unsupervised manner. Such an autoencoder consists of two parts: an *encoder*, which compresses the input to a lower-dimensional representation that captures only its most essential features, and a *decoder*, which takes this latent representation and attempts to reconstruct the original input from it. In this way, the autoencoder is able to learn and identify the underlying patterns of the clean speech signals. The CAE specifically autoencodes on the time-frequency representation of the clean speech signals, but considers

only the magnitude spectrogram and discards all phase information during training.[1] The spectrogram that was reconstructed by the decoder (referred to as $\hat{\mathbf{C}}$) is compared against the input spectrogram (referred to as $\mathbf{C}$). The cost function calculates the loss between the original and reconstructed version, and the performance of the CAE is optimized by attempting to minimize this loss during training.

2. As a second step, the clean speech and noise data are combined in order to obtain noisy mixture signals. The magnitude spectrogram representations of these mixture signals (referred to as $\mathbf{M}$) are used as input to train a mixture autoencoder (MAE). Similar to the CAE, the encoder of the MAE compresses these mixture spectrograms to a lower-dimensional representation, and the decoder attempts to reconstruct the original input from this latent representation (the reconstructed mixture spectrograms are referred to as $\hat{\mathbf{M}}$).

However, the primary goal of speech enhancement models is to learn how to reconstruct a *clean* representation from a noisy mixture, rather than reconstructing the noisy mixture itself. In order to achieve this, Wang et al. design their architecture in such a way that the MAE shares its latent space with the CAE. Please refer to Figure 4.1 for a visualization of this architectural design. Essentially, this means that the cost function used to train the MAE not only considers the reconstruction loss between the input mixture and the predicted mixture, it also compares the MAE's latent representation to the CAE's latent representations (i.e. compressed representations which capture the most important parts of the input spectrograms). As the cost function's primary objective is to minimize the difference between these representations, the MAE is encouraged to learn a mapping between the mixtures and clean representations. In Figure 4.1, this mapping is represented by the path that starts at $\mathcal{E}_{\mathbf{m}}$ (the MAE encoder) and ends at $\mathcal{D}_{\mathbf{c}}$ (the CAE decoder).

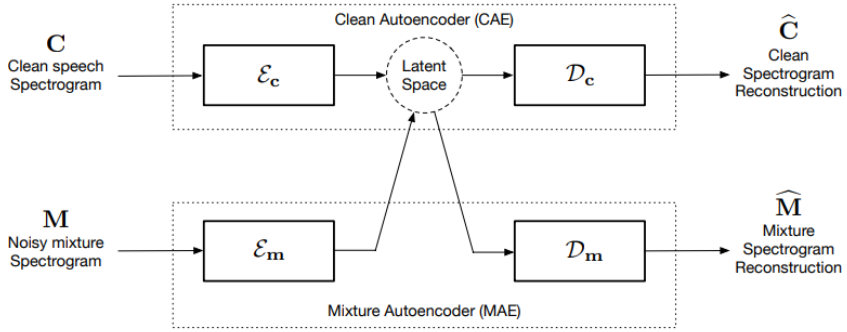


Figure 4.1: Architecture of SSE model by Wang et al. [6]

This sharing of the latent space between the CAE and MAE enables the overall model to perform speech enhancement without the need for parallel or labeled training examples

[1] Please refer back to Section 2.2 for a discussion of magnitude and phase in the time-frequency domain.

of clean and noisy speech. This is a major benefit of self-supervised learning, as parallel pairs of data can be difficult to obtain in practice. On top of that, this design makes the training process more efficient, as the two autoencoders have some design overlap and do not have to be trained completely separately.

Now that the architecture of Wang et al.'s model has been broadly discussed, it is important to elaborate on some of the finer details and parameters. First, it should be noted that both the CAE and MAE consist of a sequence of convolutional layers. In simple terms, a convolutional layer slides a small vector or matrix, also known as the kernel, over the input spectrogram to analyze it one small segment at a time. By taking the dot product of the kernel and a segment of the spectrogram, we are able to extract only the most significant features of this segment. In the encoder, this process allows us to compress or downsample the input, reducing the size of the input while retaining crucial information. In the decoder, we apply transposed convolutions instead, which reverse this process: they upsample the compressed latent representation back into the original input.

This original input spectrogram is 2-dimensional, with both a time and frequency axis. However, the convolutions which Wang et al. apply to this spectrogram are 1-dimensional. Essentially, the convolutional operation treats each frequency bin in the spectrogram as a separate "channel" and the kernel slides across the temporal dimension of such a channel, making it a 1-dimensional convolution. In the encoder of the CAE, the parameters are set such that the convolutions sequentially decrease the number of channels from 513→512→256→128→64, and this is reversed in the decoder. In the MAE, the number of channels decrease from 513→512→400→300→200→ 100→ 64 and this is once again reversed. As for the other parameters in their experimental set-up, Wang et al. set the size of the kernel to 7 and the stride to 1, meaning that the kernel contains 7 weights that determine which features within the channel are the most important, and the stride indicates the step size at which the kernel moves along the channel. The number of epochs (i.e., the number of times the model iterates through all data samples during training) was not explicitly stated in the paper. However, based on the code, the CAE appears to be trained for 700 epochs, and the MAE for 1500 epochs.

To train the CAE and MAE, Wang et al. specifically use the DAPS (Device And Produced Speech) data set [47] and the BBC noise data set [48]. However, unfortunately the BBC noise data set is not available anymore, and the speech in the DAPS data is not of the best quality. Therefore, for the baseline experiment, this research will substitute the original datasets with the LibriSpeech and UrbanSound8K sets, which will be elaborated on further in the following section.

4.2. METHODOLOGY: MODEL EXPERIMENTS

The following section will elaborate on the experiments that are conducted using Wang et al.'s model. Each of these experiments contributes to answering the main research question as stated in the introduction to this thesis, namely:

> *How effective is Wang et al.'s self-supervised speech enhancement model in removing climbing gym noise from speech signals?*

In order to address this question, it is first necessary to establish a baseline model. This

baseline serves as a benchmark for evaluating the performance of the other models that will be discussed.

MODEL 1: THE BASELINE

The primary focus of our research is to evaluate the effectiveness of Wang et al.'s self-supervised speech enhancement model in eliminating climbing gym noise from speech signals. In order to research this, it is first necessary to establish a baseline, which we call M1. M1 uses the same experimental set-up as Wang et al. and uses the same values for all described parameters. The CAE of the baseline model is trained with clean speech data derived from the LibriSpeech `train-clean-100` subset. The MAE is trained with mixtures that were created by combining clean LibriSpeech data with noise from the UrbanSound8K data set. As discussed in the previous chapter, the selected UrbanSound8K set only includes data from 4 out of 10 noise classes, as the sounds from these specific classes bear some resemblance to climbing gym noise. By doing so, we explore whether training Wang et al.'s model with noise that shares similarities to the climbing gym noise would already be adequate for effectively extracting this noise type, or whether this requires compiling a completely new training set of gym noise.

In line with Wang et al.'s approach, we train the CAE for 700 epochs and the MAE for 1500 epochs. During a single epoch, the autoencoders each process 2-second audio samples a total of 18,000 times, which is equivalent to 10 hours of training data. The CAE is trained with 10 hours of clean speech, and the MAE is trained with 10 hours of noisy mixture signals. For the training of the MAE, we ensure that all of the 4 UrbanSound8K classes are equally represented in the mixtures.

In Wang et al.'s research, these mixtures were created using two different signal-to-noise ratio (SNR) settings: 5dB and 10dB. In speech enhancement, this ratio measures the amount of clean speech as relative to the amount of background noise within a mixture, where higher values indicate less background noise. The experiments in this thesis differ slightly from Wang et al.'s research, as we decided to incorporate three different SNR settings: -5, 0 and 5. An SNR of -5 indicates that there is less speech than noise in the signal, an SNR of 0 indicates equal amounts of speech and noise, and an SNR of 5 indicates that there is more speech than noise (or, to be specific: the speech signal level is 5 decibels higher than the noise level). Training the model with mixtures created at these three different SNR settings should improve the model's ability to generalize and perform accurately in diverse and noisy audio environments.

After training M1, we test the performance of this model on two types of mixtures: the type it was trained with, i.e., LibriSpeech-UrbanSound8K mixtures, and a different set of mixtures consisting of clean LibriSpeech data combined with climbing gym noise. By conducting this comparative analysis, we can observe how efficient the model already is at extracting climbing gym noise from speech signals without having encountered this exact noise during training.

MODEL 2: ADDITIONAL TRAINING DATA

In the next experiment, we add some of the collected gym noise to the UrbanSound8K training data set. This model is referred to as M2. Once again, each of the UrbanSound8K noise classes, now with the added climbing gym noise class, are equally represented in

the overall training set of mixtures. To ensure a fair comparison with the baseline model, we maintain the same experimental set-up, keeping the number of epochs and other relevant parameters consistent.

After training, M2 is tested on mixtures of LibriSpeech data and climbing gym noise, in order to gauge whether adding this noise type to the training samples will improve the effectiveness of Wang et al.'s model in removing climbing gym noise from speech signals. Given that the baseline results have already been established in the previous experiment, there is no need to retest on the LibriSpeech-UrbanSound8K mixtures. Please refer to Figure 4.2 for an overview of the M1 and M2 experiments.

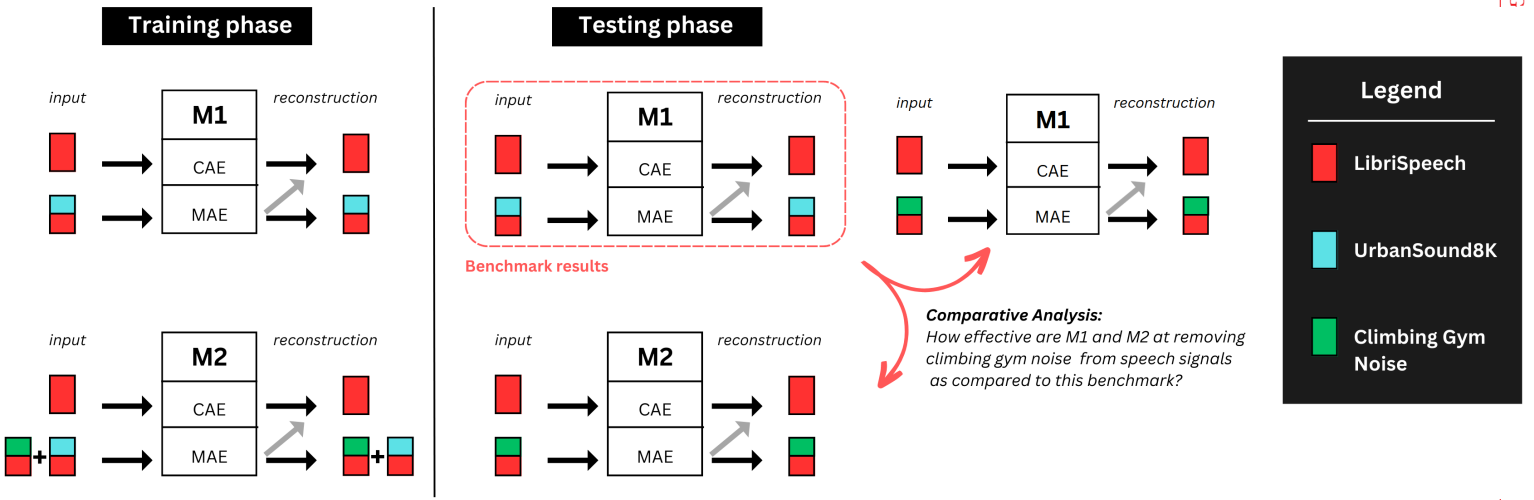


Figure 4.2: Overview of M1 and M2 experiments

MODEL 3: MODIFIED ARCHITECTURE

In the third model experiment, called M3, we make several modifications to the architecture of Wang et al.'s model. As mentioned before, their model only processes the magnitude spectrogram of the input signal, which is obtained by squaring the magnitude information of the Short-Time Fourier Transform (STFT). The phase spectrogram, which is crucial for understanding the phase relationships between different frequency components of the speech signal, can be obtained by computing the angle of each element in the STFT matrix. Despite the importance of this information, the phase spectrogram is discarded during training due to its highly unstructured character. However, as mentioned in Chapter 2, neglecting this information can impose an upper bound on the performance of the speech enhancement model [10].

This is why, for M3, we propose an alternative architectural design. As previously explained in Section 2.2 of Chapter 2, the complex-valued STFT of a signal can be expressed in multiple ways: the complex values can be represented in terms of their polar coordinates (the magnitude and phase angle), or using Cartesian coordinates with a complex exponential, where the complex values are expressed in terms of their real and imaginary parts. In M3, we opt for this latter approach, and compute the real and imaginary spectrograms from these values, which both exhibit a clear structure. Since both the real and imaginary spectrograms contain magnitude and phase information, this approach ensures that the phase information is retained.

Whereas Wang et al.'s model simply processed the magnitude spectrogram, M3 has to process both the real and imaginary spectrograms. The original 2-dimensional input now becomes 3-dimensional instead: we stack these two spectrograms together to form a 3D "image" to feed as input to the network. The reason for stacking the spectrograms together is that it enables the model to capture more complex patterns and effectively utilize both magnitude and phase information. Please refer to Figure 4.3 for a visualization of this change.

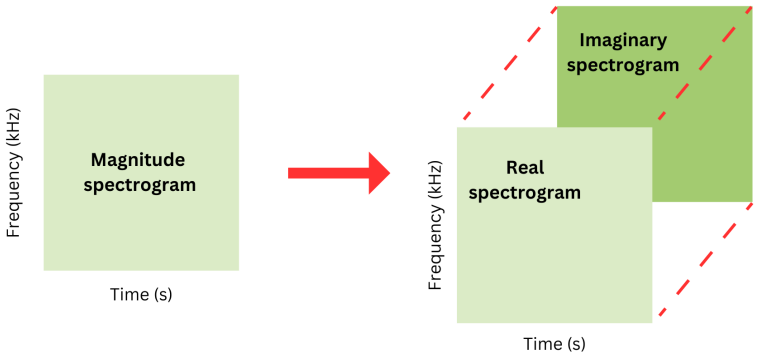


Figure 4.3: From 2D to 3D input

Due to the additional dimension in the input, M3 also employs 2D rather than 1D convolutions. The original 1D convolutions only considered the length of each channel in the

magnitude spectrogram. In contrast, the 2D convolutions take into account the channels in both the real and imaginary spectrogram. This enables the model to capture spatial relationships in the data more effectively.

In an ideal scenario, we would like all layers in both the CAE and MAE to process 3D data. However, doing so would significantly increase the model size and processing latency. Therefore, it is necessary to strike a balance between preserving as much information about the input signal as possible, and ensuring the model remains compact and fast. To achieve this balance, we adopt a specific approach for both the CAE and MAE. For each autoencoder, we design the encoder so that only its *first* layer handles 3D input, and the decoder's *last* layer generates 3D output. In all the intermediary layers, the 3D data is collapsed into a 2D representation. See Figure 4.4 for a visualization of the described adjustment.

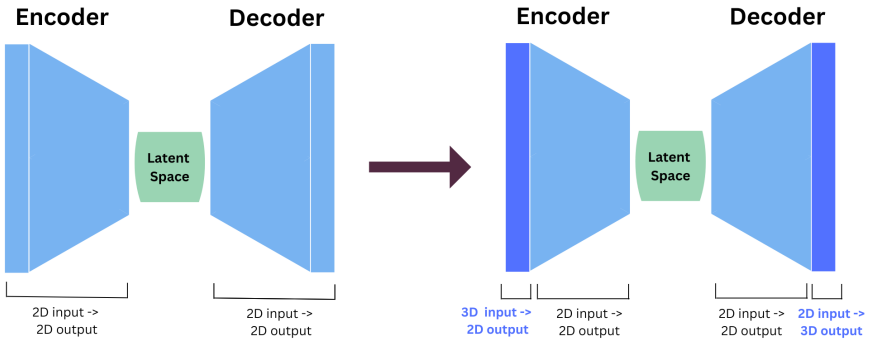


Figure 4.4: Autoencoder with 3D input and output

Incorporating these modifications in M3 allows us to explore the impact of retaining phase information on the effectiveness of Wang et al.'s model. After training the model with this modified approach, we evaluate its performance on mixture signals consisting of clean LibriSpeech data and climbing gym noise. By comparing the performance of the modified M3 to that of M2, we establish whether retaining phase information in this way significantly increases the model's effectiveness at extracting climbing gym noise from speech signals.

4.2.1. DEMONSTRATOR

Wang et al.'s [6] code for the original model, which was used to run M1 and M2, can be found here. [2]. For M3, we have adapted this code and published it as a separate GitHub project. The demonstrator of the M3 experiment can be found at this GitHub project.[3]

[2] https://github.com/jeffreyjeffreywang/SSE
[3] https://github.com/ehwgal/SSE-modified

Most of the adjustments implemented can be found in the `model.py` file. Instructions on how to run the demonstrator can be found in the `README.md` file of this project.

4.3. EVALUATION METRICS

Now that the model experiments have been broadly discussed, it is important to elaborate on how these models will be evaluated. We use five different evaluation metrics for this purpose, which can be observed in Table 4.1.

evaluation metric	explanation of abbreviation	intrusive vs. non-intrusive	evaluation target: quality vs. intelligibility	scoring range (bad - good)
CBAK	predictor of background intrusiveness	intrusive	perceptual quality	1 to 5 [4]
NISQA	speech quality and naturalness assessment	non-intrusive	perceptual quality	1 to 5
PESQ	perceptual evaluation of speech quality	intrusive	perceptual quality	1 to 5
SSNR	segmental signal-to-noise ratio	intrusive	intelligibility	-10 to 35
STOI	short-time objective intelligibility	intrusive	intelligibility	0 to 1

Table 4.1: Evaluation metrics for assessing model performance

These five metrics were chosen as they evaluate different aspects of the predicted speech signal (also known as the *enhanced* signal). CBAK [49], NISQA [50], and PESQ [51] are all objective algorithms that were developed through extensive statistical analysis of MOS (Mean Opinion Score) tests. MOS tests are subjective listening tests in which human annotators rate the quality of an audio signal from 1 (bad) to 5 (good). By employing different statistical techniques to uncover patterns in this data, the algorithms are developed to have their scores correlate with the subjective scores gathered in these tests. These algorithms are often used in speech enhancement tasks and are meant to forego the need for human evaluators.

SSNR [52] and STOI [53], on the other hand, are employed to evaluate the intelligibility of the enhanced speech signal. SSNR is an objective evaluation metric that first divides the audio signal into shorter segments or frames. Only when speech is identified in a frame, it estimates the ratio of the speech signal power as compared to the noise power. It does this for both the original clean speech signal as well as the enhanced speech signal that was predicted from the mixture. By assessing how well the enhanced signal resembles the original clean speech signal, the SSNR is able to measure how effective the speech enhancement model is at denoising the mixture signal. In theory, the SSNR can range from negative infinity to infinity. However, for practical purposes, this metric is usually kept within a certain range – which, in our case, is -10 to 35dB.

STOI is another metric used for evaluating intelligibility. This score also divides the signal into shorter overlapping frames and compares the clean speech signal to the en-

[4] Although CBAK, NISQA and PESQ technically maintain a scale from 1 to 5, it is important to note that some versions of these algorithms use a scale from -0.5 to 4.5 for mathematical convenience.

hanced or predicted speech signal. Its score ranges from 0 (no intelligibility) to 1 (perfect intelligibility). One important difference is that STOI normalizes and "clips" the (time-frequency representation of) the audio signals before comparison. Clipping means that STOI sets an upper bound to how severely the audio can be degraded, which ensures the metric is less sensitive to extreme outliers and leads to a more robust evaluation.

It is important to note that out of the five metrics, NISQA is the only *non-intrusive* evaluation metric. The difference is that the intrusive metrics require comparison material: they evaluate the enhanced signal in comparison to the original clean speech signal provided to them. NISQA, on the other hand, is able to evaluate the enhanced signal without this reference: the neural network is trained on expert MOS scores, and its trained model weights can be used to predict the quality of a speech signal. It should be noted that since NISQA makes this prediction without access to a clean reference, there is a potential risk that the data sets used for training NISQA are too dissimilar from the data used in our own model experiments. This shows why it is necessary to employ multiple evaluation metrics.

Now that the methodology for evaluation and the model experiments have been thoroughly discussed, the next chapter will provide a detailed overview of the results obtained from the experiments.

5

RESULTS

This chapter first briefly discusses how to read and interpret the results tables, before providing an objective analysis of the results per model experiment. An interpretation of these outcomes will be provided in the next chapter.

5.1. PRESENTATION OF RESULTS

In this section, we will shortly discuss how to read the results presented in each table. Table 5.1 presents the results of the M1 model tested on the LibriSpeech-UrbanSound8K mixtures. During testing, we assess the model performance individually for each of the UrbanSound8K noise classes at different SNR settings (-5, 0, 5) to ensure a more comprehensive evaluation. Specifically, for each combination of noise class and SNR setting, we create 20 mixture signals in total that each contain one full sentence by a single speaker. The table's vertical axis displays the individual noise classes and SNR settings, while the horizontal axis shows the evaluation metrics together with their scoring ranges. Each metric includes results for both the original mixture and the enhanced speech signal that was predicted by the model.

Table 5.2 is structured in the same way and shows the results of the M1 model on mixtures composed of LibriSpeech and climbing gym noise. For a more thorough evaluation, we have separated the climbing gym noise into distinct classes that are meant to parallel the UrbanSound8K classes. However, as none of the recordings appeared to feature the sounds of an air conditioner very prominently (which is class 0 in the UrbanSound8K data), the first class is a more general class that includes a diverse range of climbing gym environment noises. The rest of the classes are meant to resemble the UrbanSound8K data: the second class comprises only noise from children's group activities in the climbing hall, encompassing sounds like children's voices and climbing activities in the clip 'n climb area. The third class contains noises from route setting activities, which involves the installation of new climbing routes on the walls and may include drilling noises. Finally, the fourth class only contains noise from music being played loudly in the gym. By aligning the noise classes in the two data sets in this way, we can directly compare M1's results and its adaptability to different types of noise.

Noise class	SNR (dB)	CBAK (1 to 5)		PESQ (1 to 5)		SSNR (-10 to 35)		STOI (0 to 1)		NISQA (1 to 5)	
		mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced
class 0 *air conditioner*	**-5**	1.39	1.45	1.05	1.10	- 4.52	- 2.63	0.51	0.52	2.24	2.32
	0	1.56	1.63	1.06	1.14	- 3.07	- 1.03	0.60	0.63	2.22	2.16
	5	1.87	1.82	1.15	1.26	- 0.31	0.33	0.70	0.70	2.14	2.18
class 2 *children playing*	**-5**	1.38	1.36	1.10	1.07	- 3.85	- 2.35	0.48	0.47	1.94	2.14
	0	1.66	1.54	1.18	1.12	- 2.21	- 1.47	0.57	0.57	2.12	2.14
	5	1.83	1.78	1.14	1.21	- 0.16	0.35	0.66	0.67	2.21	2.19
class 4 *drilling noises*	**-5**	1.49	1.52	1.08	1.07	- 4.53	- 2.41	0.47	0.48	1.95	1.91
	0	1.62	1.64	1.06	1.10	- 2.73	- 0.94	0.56	0.59	2.07	1.77
	5	1.91	1.87	1.11	1.18	- 0.19	0.89	0.69	0.70	2.19	2.06
class 9 *street music*	**-5**	1.32	1.35	1.12	1.10	- 4.33	- 2.82	0.46	0.47	2.11	2.25
	0	1.61	1.58	1.12	1.15	- 2.45	- 1.45	0.58	0.60	2.09	2.14
	5	1.75	1.72	1.15	1.23	- 0.11	0.03	0.70	0.70	2.10	2.31

Table 5.1: M1 results on UrbanSound8K test set

Noise class	SNR (dB)	CBAK (1 to 5)		PESQ (1 to 5)		SSNR (-10 to 35)		STOI (0 to 1)		NISQA (1 to 5)	
		mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced
all gym noise	**-5**	1.53	1.61	1.07	1.13	- 3.15	- 0.96	0.64	0.63	2.12	2.30
	0	1.75	1.79	1.12	1.23	- 1.18	0.30	0.74	0.71	2.27	2.26
	5	2.07	1.97	1.24	1.36	1.67	1.48	0.82	0.76	2.40	2.21
children climbing	**-5**	1.46	1.60	1.11	1.12	- 3.74	- 1.11	0.61	0.61	2.21	2.30
	0	1.66	1.81	1.11	1.23	- 1.94	0.38	0.71	0.70	2.48	2.27
	5	1.95	2.01	1.19	1.37	0.75	1.81	0.80	0.77	2.52	2.25
route setting	**-5**	1.54	1.58	1.08	1.10	- 3.81	- 1.86	0.56	0.53	2.21	2.19
	0	1.73	1.74	1.12	1.17	- 2.19	- 0.55	0.65	0.62	2.32	2.15
	5	2.00	1.91	1.22	1.26	0.29	0.83	0.73	0.70	2.32	2.13
music	**-5**	1.41	1.46	1.06	1.10	- 3.42	- 1.85	0.57	0.57	2.34	2.29
	0	1.64	1.67	1.10	1.17	- 1.50	- 0.42	0.68	0.68	2.32	2.27
	5	1.95	1.89	1.19	1.30	1.24	1.08	0.78	0.75	2.27	2.20

Table 5.2: M1 results on climbing gym noise test set

Following the same format as Table 5.2, Table 5.3 showcases the outcomes of the M2 model on the climbing gym noise mixtures, and Table 5.4 displays the results of the M3 model on these same mixtures.

Having covered the presentation of the results tables, the next sections will provide a brief analysis of the model outcomes.

5.2. M1 RESULTS

Table 5.1 presents the results of M1 on the LibriSpeech-UrbanSound8K mixtures (the baseline), and Table 5.2 shows the results of this model on the LibriSpeech-climbing gym noise mixtures. One interesting observation to make is that the mixtures created with the climbing gym noise set generally appear to score higher than mixtures created with UrbanSound8K noise. However, for both tables, it can be seen that there is generally very little to no improvement between the scores for the original mixtures and the enhanced speech signals that were denoised by the model. Occasionally, the enhanced signal scores even fall below the original mixture scores.

Notably, in Table 5.1, SSNR and STOI (which measure intelligibility) seem to be the only metrics where the enhanced signal consistently matches or surpasses the mixture's scores. In contrast, in Table 5.2, metrics like CBAK, PESQ, and SSNR tend to improve, except in some cases where the SNR is set to 5 – meaning that the background noise level in the mixture is already 5dB lower than the clean speech level – whereas STOI and NISQA tend to decrease more often. However, it should be noted that the score differences in both tables are so small as to be almost negligible. It is therefore difficult to observe clear trends within the results.

5.3. M2 RESULTS

Please refer to Table 5.3 for the results of the M2 experiment on the gym noise mixtures. In M2, we introduced an additional class of climbing gym noise into the training data and evaluated its performance on mixtures containing climbing gym noise. We then compare these findings with the results of M1 tested on climbing gym noise (Table 5.2), in order to gauge whether the inclusion of gym noise in the training data improves the model's ability to remove climbing gym noise from speech signals.

When we compare the results of M2 with those of M1, a noticeable pattern emerges: the enhanced signal scores for nearly all metrics show a fairly slight increase, or at least remain consistent, when compared to the first model experiment. The scores have minimally decreased only in the case of the music noise class for the CBAK metric. However, the major exception is the NISQA metric, for which the majority of scores for the enhanced signals have decreased as compared to the first model experiment. A potential explanation of why this might be will be provided in the next chapter.

Noise class	SNR (dB)	CBAK (1 to 5)		PESQ (1 to 5)		SSNR (-10 to 35)		STOI (0 to 1)		NISQA (1 to 5)	
		mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced
all gym noise	-5	1.53	1.61	1.07	1.14	- 3.15	- 0.83	0.64	0.64	2.12	2.17
	0	1.75	1.80	1.12	1.25	- 1.18	0.46	0.74	0.72	2.27	2.17
	5	2.07	1.98	1.24	1.38	1.67	1.64	0.82	0.77	2.40	2.14
children climbing	-5	1.46	1.61	1.11	1.14	- 3.74	- 0.86	0.61	0.62	2.21	2.16
	0	1.66	1.82	1.11	1.25	- 1.94	0.71	0.71	0.71	2.48	2.13
	5	1.95	2.03	1.19	1.39	0.75	2.15	0.80	0.77	2.52	2.17
route setting	-5	1.54	1.58	1.08	1.10	- 3.81	- 1.57	0.56	0.55	2.21	2.21
	0	1.73	1.75	1.12	1.18	- 2.19	- 0.20	0.65	0.64	2.32	2.20
	5	2.00	1.92	1.22	1.28	0.29	1.17	0.73	0.73	2.32	2.17
music	-5	1.41	1.45	1.06	1.10	- 3.42	- 1.85	0.57	0.58	2.34	2.17
	0	1.64	1.66	1.10	1.18	- 1.50	- 0.40	0.68	0.68	2.32	2.17
	5	1.95	1.88	1.19	1.30	1.24	1.10	0.78	0.75	2.27	2.17

Table 5.3: M2 results on climbing gym noise test set

Noise class	SNR (dB)	CBAK (1 to 5)		PESQ (1 to 5)		SSNR (-10 to 35)		STOI (0 to 1)		NISQA (1 to 5)	
		mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced	mixture	enhanced
all gym noise	-5	1.53	1.08	1.07	1.19	- 3.15	-10.0	0.64	0.29	2.12	1.28
	0	1.75	1.06	1.12	1.11	- 1.18	-10.0	0.74	0.29	2.27	1.39
	5	2.07	1.03	1.24	1.06	1.67	-10.0	0.82	0.29	2.40	1.31
children climbing	-5	1.46	1.04	1.11	1.13	- 3.74	-10.0	0.61	0.29	2.21	1.32
	0	1.66	1.03	1.11	1.08	- 1.94	-10.0	0.71	0.29	2.48	1.37
	5	1.95	1.02	1.19	1.05	0.75	-10.0	0.80	0.29	2.52	1.31
route setting	-5	1.54	1.11	1.08	1.08	- 3.81	-10.0	0.56	0.29	2.21	1.29
	0	1.73	1.06	1.12	1.12	- 2.19	-10.0	0.65	0.29	2.32	1.30
	5	2.00	1.07	1.22	1.22	0.29	-10.0	0.73	0.29	2.32	1.24
music	-5	1.41	1.03	1.06	1.07	- 3.42	-10.0	0.57	0.29	2.34	1.27
	0	1.64	1.02	1.10	1.06	- 1.50	-10.0	0.68	0.29	2.32	1.24
	5	1.95	1.03	1.19	1.08	1.24	-10.0	0.78	0.29	2.27	1.37

Table 5.4: M3 results on climbing gym noise test set

5.4. M3 RESULTS

Please refer to Table 5.4 for the results of the M3 experiment on the gym noise mixtures. For this model, we experimented with incorporating phase information during the training of the model, in order to gauge whether the inclusion of this information could enhance the model's effectiveness in eliminating gym noise from speech signals. As can be seen in this table, M3 scores significantly worse than any of the previous model experiments. This is especially apparent for the SSNR and STOI metrics, which are the two metrics measuring intelligibility. Both these metrics seem to get stuck at very low scores and show no improvement across different noise classes or SNR settings. In the upcoming chapter, we will discuss potential factors that could have contributed to this outcome.

Now that the results have been presented and briefly analyzed, the next chapter will provide an interpretation of the obtained results and discuss their relevance in the context of the primary research question of this thesis.

6

Discussion and Conclusion

This chapter will provide an interpretation of the results and discuss them in light of the research question and hypothesis of this thesis. It will also suggest potential directions for future research and discuss any challenges encountered during this study.

6.1. Discussion

The main aim of this thesis was to investigate the effectiveness of Wang et al.'s self-supervised speech enhancement model in removing climbing noise from speech signals. As stated in Section 1.1, we hypothesized that the unique noise characteristics of climbing gym noise, as well as the exclusion of phase information during model training, could potentially limit the model's performance. We additionally hypothesized that incorporating the phase during training could improve the model's denoising capabilities.

To investigate these hypotheses, we conducted three separate model experiments. The first model experiment, M1, was trained on a slightly different noise dataset (UrbanSound8K), and evaluated on mixtures created with this noise to establish a baseline. Subsequently, this model was evaluated on mixtures composed of clean speech and climbing gym noise. This was done in order to evaluate how efficient the model is at extracting climbing gym noise from speech signals without having encountered this exact noise type during training. In the second experiment, M2, we added climbing gym noise to the training data in order to see whether this would make the model more effective. Finally, in M3, we adapted the architecture of the original model to incorporate phase information, with the expectation that this would improve model performance.

After analyzing the results for the M1 and M2 experiments, it became evident that the difference between the scores for the original mixtures and the enhanced speech signals was remarkably small, meaning that the models are underperforming. Even when M1 was evaluated on mixtures containing UrbanSound8K noise, i.e. the same data set it was trained on, the model scored significantly low. This showcases that the model's effectiveness was limited in both experiments. However, this outcome might have been partly due to our experimental set-up. Although the different UrbanSound8K classes

were combined into one data set to reflect the diversity of the climbing gym noise, these individual noise classes could potentially be too distinct from each other, and the training data for each noise class might have been too limited. This potential disparity among the noise classes could have prevented the model from learning clear noise patterns during the training phase.

Although we observed that M2 performed slightly better on the climbing gym noise test set than M1, the increase was minimal (usually only one or two decimals) and the improvement seems modest given the amount of gym noise training data that was added. Once again, the potential disparity between the UrbanSound8K classes could have hindered the model in identifying clear noise patterns. Had there been more time, it would be worth experimenting with other data and larger portions of climbing gym noise. During the M2 experiment, we also observed that the mixtures created with climbing gym noise generally scored higher than those created with UrbanSound8K noise. One potential reason for this trend could be the uncurated character of the climbing gym noise data set. Some of the noise recordings in this set were relatively quiet due to the distance between the microphone and noise sources. Although this variation in noise levels accurately reflects the fluctuating volume levels of different noise sources within a real climbing gym environment, it would have perhaps been better to categorize and evaluate the gym noise recordings based on their volume levels.

Another notable finding from the M2 experiment is that the scores for all metrics improved except for NISQA. The most probable explanation for this pattern is the non-intrusive character of the NISQA metric. Unlike the other metrics, NISQA does not rely on a clean reference signal for comparison – instead, it utilizes trained model weights to predict a score for the enhanced signal independently. Since NISQA was trained with audio from different acoustic environments, there is a chance that this metric might not be well suited to our specific use case. This hypothesis was confirmed through further experimentation, which revealed that NISQA did not even award the clean samples very high scores – in fact, none of the clean recordings surpassed a score of 2.50. As a direct consequence, this means that the enhanced signals would also never be able to obtain a score higher than 2.50 for the NISQA metric, so it is no wonder these scores remained low.

Finally, the M3 experiment was conducted to gauge whether retaining phase information during model training would improve results. Although the model runs without errors, the poor results show that further refinement of the model architecture is necessary. While the perceptual quality metrics (CBAK, PESQ, NISQA) generated different scores for each scenario, the intelligibility metrics (SSNR and STOI) seemed to get stuck at the same low scores across all test runs. Since the inclusion of phase information is specifically intended to improve the intelligibility of the predicted signal, this observation strongly implies that the model still struggles to capture phase relationships effectively. One reason for this could be that the collapse from the 3D representations into the 2D representations is too rigorous, and the model is unable to capture the spatial relationships well.

Based on the discussed findings, it is difficult to provide a clear answer to the research question, and definitively validate or reject the hypothesis of this research. Although the model seems to underperform in all cases and not remove climbing gym noise effec-

tively, this could have been influenced by our experimental set-up and choice of data. Furthermore, the M3 model's inability to effectively capture phase relationships shows that further research needs to be conducted. The following section provides possible directions for future research that would allow us to provide a more definitive answer to the research question.

6.2. RECOMMENDATIONS FOR FUTURE RESEARCH

There are several approaches that can be taken to improve on the current research. First of all, it is clear that the model is not able to identify the noise patterns effectively, which could have been influenced by our choice of data. Experimenting with more focused training on the individual noise classes, as well as creating larger and more diverse training sets for each of these classes, could provide valuable insights into how the model responds to the different noise characteristics. This should address whether the poor results are a consequence of the dissimilarity between noise classes and the limited size of these classes.

Furthermore, although not implemented in this study, Wang et al. [6] conducted some additional experiments where they trained their model with different amounts of "pure noise" samples. This means that the model was exposed to a certain portion of training samples that contained only background noise, in order to have it learn the noise patterns more effectively. However, although the performance of the model slightly increased, the improvement in results did not seem significant enough to warrant following these experiments in this thesis.

Second of all, it is clear that the NISQA metric is not well suited to our use case. It would therefore be a good idea to either choose data which is more compatible with NISQA, or to incorporate other non-intrusive metrics that might perform more effectively for our use case. An example that could be explored is STOI-Net [54], which is the non-intrusive version of the STOI metric already employed. Since STOI and STOI-Net essentially evaluate the same aspects of the enhanced signal, this would also allow for a rather interesting comparison between the scores for these metrics. Of course, it should be mentioned that gathering feedback from human evaluators next to the employed algorithmic metrics would be a very valuable contribution to the evaluation part of this research. However, as expert human evaluation can be expensive to collect and can take up a significant amount of time, most speech enhancement studies tend to confine themselves to algorithm-based metrics only.

Finally, the observed findings showcased that the modified model, M3, appears not to be able to capture the phase relationships within the input very effectively. As discussed, one potential reason for this could be that the collapse from the 3D representations into the 2D representations is too rigorous, and the model is unable to capture the spatial relationships well. There are several approaches that could be taken to have the model process the dimensionality of the data more efficiently. The most obvious suggestion would be to not collapse the 3D input into a 2D representation at all, but to have all layers of the model process 3D data. Unfortunately, this would significantly increase the model size and processing latency. Another option, however, could be to experiment with *residual blocks*, which have the potential to improve the spatial information flow through the network.

Residual blocks were first introduced in a 2016 paper by He et al. [55]. This approach can be used to tackle the problem of *vanishing gradients*, which can hinder the training of deep neural networks. In traditional neural networks, the input information is processed and transformed through the layers in a sequential order to obtain the corresponding output. The difference between this predicted output and the desired output is calculated by a loss function. In order to improve the predictions of the model, gradient values are computed through backward propragation across the network's layers. These gradients indicate how much each weight in the network should be adjusted in order to minimize the calculated loss. However, if the network contains a lot of layers and becomes very complex, meaning that the input and output are far removed from each other, we might encounter the problem of vanishing gradients. This means that the gradients values become so small that the network is not able to effectively update its weights, and often stops learning from the data altogether.

Residual blocks can be used to counter this problem. Essentially, this approach entails that the layers of the neural network are divided into smaller blocks of layers, and within these blocks information flows both through and around the layers, by means of so-called *skip connections*. These skip connections allow the input information to bypass certain layers and directly flow from the input to the output of such a block. This kind of approach ensures that less information is lost along the way, and the gradient values are more likely to be preserved during backpropagation, leading to improved learning of the model.

Given that our network seems to not capture the spatial relationships in the 3D input data very effectively, it would be sensible to experiment with residual blocks to determine whether they contribute to the preservation of this information. If residual blocks are not an option, another idea would be to experiment with different activation functions, which have a significant influence on the learning process of the network. Whereas the current network works utilizes a softplus activation in all its convolutional layers, it might be interesting to try out other variants like ReLu or LeakyReLu.

6.3. CHALLENGES

Having discussed the outcomes of this study and proposed possible directions for future investigation, some of the challenges which were encountered during this research will now be discussed. First of all, the code implementation was made difficult by the lack of a proper `requirements.txt` file and by the fact that several of the functions were quite poorly documented. This meant that numerous dependency conflicts had to be resolved before the code was able to run. Furthermore, as the code was written for a specific data set which is currently not available anymore (the BBC.16k [48] data), it was at first difficult to ascertain how the noise files were loaded for model training.

Once we had established a `requirements.txt` file and resolved all dependency conflicts, it turned out that the PyTorch and CUDA versions were not compatible with the HPC cluster of the University of Groningen. Unfortunately, it was not clear at first that this was the issue as the model returned no error messages. It therefore took quite a long time to fix this problem. The HPC cluster also experienced GPU issues several times, which meant we were temporarily not able to train the model.

When the cluster was working, it often took some time for the GPU node to become

available. Although this is to be expected as we requested very long jobs for model training, the combined waiting and training time of each experiment took 2 days minimum. This meant that we had to wait ca. 2 days each time we adjusted an architectural component to see if training had been successful. In future experiments, the training time can be significantly reduced by decreasing the number of epochs (see Appendix B) and getting rid of the metric calculations in between epochs in the training phase.

Lastly, it turned out that the code we used lacked a proper evaluation script. While the model processes 2-second audio samples, the evaluation phase required testing on complete sentences of audio. As a result, we spent a considerable amount of time developing a testing script which was able to accurately assess model performance.

6.4. CONCLUSION

This study set out to investigate how effective Wang et al.'s [6] self-supervised speech enhancement model is at removing climbing gym noise from speech signals. This was investigated through means of several different model experiments, which all examined different factors that could potentially have an effect on the model's denoising capabilities. Specifically, these experiments directly tested our hypotheses, in which we speculated that the unique characteristics of climbing gym noise and the exclusion of phase information during training could significantly hinder the performance of the model.

While the experiments did not provide a definitive answer to the primary research question and we were not able to reject or validate the hypotheses, they did provide valuable insights for potential future research. None of the experiments yielded satisfactory results, but this could have been partly due to dissimilarity among noise classes and limitations of the training data set. Further training of the M1 model with more targeted noise data could shed light on the way the model is able to learn noise patterns, and could potentially improve its denoising capabilities.

Similarly, for M2, the experiment showed that adding climbing gym noise to the training data lead to only a slight improvement, which suggests that a more refined curation of the gym noise data set might be beneficial. The experiments also highlighted the complexity of evaluating speech enhancement tasks, as the NISQA metric proved not to be well-suited to our use case, which underscores the need for human evaluation beside algorithmic metrics to gain a better insight into model performance.

While previous research has dicussed the importance of phase information for speech enhancement performance, the M3 model evidently struggled to capture the phase relationships in the input signals effectively, showcasing that the architecture needs to be refined further. Having all layers of the model process 3D data, or experimenting with residual blocks and alternative activation functions could potentially enhance the model's ability to preserve spatial information.

Although it is difficult to provide a clear answer to the research question with the discussed model outcomes, this thesis can serve as a foundation to build on for future research. By addressing the limitations of this study, researchers will be able to gain more insight into complex noise removal in real-world environments such as climbing gyms, and will be able to develop more effective models for improving speech communication in challenging noisy conditions.

BIBLIOGRAPHY

[1] Hans-Günter Hirsch and David Pearce. “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”. In: *ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*. 2000.

[2] Jon Barker et al. “The fifth ‘CHiME’ speech separation and recognition challenge: dataset, task and baselines”. In: *arXiv preprint arXiv:1803.10609* (2018).

[3] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. “The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings”. In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. AIP Publishing. 2013.

[4] YJSC Hu. “Subjective evaluation and comparison of speech enhancement algorithms”. In: *Speech Communication* 49 (2007), pp. 588–601.

[5] Elja Leijenhorst. “Fine-tuning ASR to specific acoustic environments: noise robustness in a climbing gym”. MA thesis. University of Groningen, 2023.

[6] Yu-Che Wang, Shrikant Venkataramani, and Paris Smaragdis. “Self-supervised learning for speech enhancement”. In: *arXiv preprint arXiv:2006.10388* (2020).

[7] Hyeong-Seok Choi et al. “Phase-aware speech enhancement with Deep Complex U-Net”. In: *International Conference on Learning Representations*. 2019.

[8] Pejman Mowlaee and Rahim Saeidi. “Time-frequency constraints for phase estimation in single-channel speech enhancement”. In: *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 2014, pp. 337–341.

[9] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. “The importance of phase in speech enhancement”. In: *speech communication* 53.4 (2011), pp. 465–494.

[10] Chuanxin Tang et al. “Joint time-frequency and time domain learning for speech enhancement”. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 3816–3822.

[11] Daniel Michelsanti et al. “An overview of deep-learning-based audio-visual speech enhancement and separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1368–1396.

[12] Jacob Benesty. *Fundamentals of speech enhancement*. Springer, 2018.

[13] Nasir Saleem, Muhammad Irfan Khattak, and Elena Verdú. “On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms”. In: *International Journal of Interactive Multimedia and Artificial Intelligence* (2020).

[14] Nabanita Das et al. “Fundamentals, present and future perspectives of speech enhancement”. In: *International Journal of Speech Technology* 24 (2021), pp. 883–901.

[15] Zili Huang et al. “Investigating self-supervised learning for speech enhancement and separation”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6837–6841.

[16] Zhifeng Kong et al. “Speech denoising in the waveform domain with self-attention”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7867–7871.

[17] Soha A Nossier et al. “A Comparative Study of Time and Frequency Domain Approaches to Deep Learning based Speech Enhancement”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[18] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[19] Arian Azarang and Nasser Kehtarnavaz. “A review of multi-objective deep learning speech denoising methods”. In: *Speech Communication* 122 (2020), pp. 1–10.

[20] Jean-Marc Valin. “A hybrid DSP/deep learning approach to real-time full-band speech enhancement”. In: *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE. 2018, pp. 1–5.

[21] Dmitrii Mukhutdinov et al. “Deep learning models for single-channel speech enhancement on drones”. In: *IEEE Access* 11 (2023), pp. 22993–23007.

[22] Dacheng Yin et al. *PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network*. 2019. arXiv: 1911.04697 [cs.SD].

[23] Electronics Notes. *What is a sine wave - electronics waveform*. URL: https://www.electronics-notes.com/articles/basic_concepts/electronic-electrical-waveforms/sine-waveform.php.

[24] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. “Complex ratio masking for monaural speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 24.3 (2016), pp. 483–492.

[25] Yusheng Tian, Wei Liu, and Tan Lee. “Diffusion-Based Mel-Spectrogram Enhancement for Personalized Speech Synthesis with Found Data”. In: *arXiv preprint arXiv:2305.10891* (2023).

[26] Md Ekramul Hamid et al. “Single channel speech enhancement using adaptive soft-thresholding with bivariate EMD”. In: *International Scholarly Research Notices* 2013 (2013).

[27] Dequan Wang and Jae Lim. “The unimportance of phase in speech enhancement”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30.4 (1982), pp. 679–681.

[28] Naijun Zheng and Xiao-Lei Zhang. “Phase-aware speech enhancement based on deep neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2018), pp. 63–76.

[29] *Complex numbers - How to Think Like a Computer Scientist - C++*. URL: https://runestone.academy/ns/books/published/thinkcpp/Chapter14/complex_numbers.html.

[30] Yanxin Hu et al. *DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement*. 2020. arXiv: 2008.00264 [eess.AS].

[31] Lars Hertel, Huy Phan, and Alfred Mertins. "Comparing time and frequency domain for audio event recognition using deep learning". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2016, pp. 3407–3411.

[32] Ashutosh Pandey and DeLiang Wang. "A new framework for CNN-based speech enhancement in the time domain". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (2019), pp. 1179–1188.

[33] Santiago Pascual, Antonio Bonafonte, and Joan Serra. "SEGAN: Speech Enhancement Generative Adversarial Network". In: *arXiv preprint arXiv:1703.09452* (2017).

[34] Yi Luo and Nima Mesgarani. "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[35] Ashutosh Pandey and DeLiang Wang. "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6875–6879.

[36] Xiaoyu Bie et al. "Unsupervised speech enhancement using dynamical variational autoencoders". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2993–3007.

[37] Xiaoyu Lin et al. "Unsupervised speech enhancement with deep dynamical generative speech and noise models". In: *arXiv preprint arXiv:2306.07820* (2023).

[38] Madhav Mahesh Kashyap et al. "Speech denoising without clean training data: A noise2noise approach". In: *arXiv preprint arXiv:2104.03838* (2021).

[39] Mostafa Sadeghi and Romain Serizel. "Fast and efficient speech enhancement with variational autoencoders". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[40] Bilal Dendani, Halima Bahi, and Toufik Sari. "Self-Supervised Speech Enhancement for Arabic Speech Recognition in Real-World Environments." In: *Traitement du Signal* 38.2 (2021).

[41] Vassil Panayotov et al. "Librispeech: an asr corpus based on public domain audio books". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2015, pp. 5206–5210.

[42] J. Salamon, C. Jacoby, and J. P. Bello. "A Dataset and Taxonomy for Urban Sound Research". In: *22nd ACM International Conference on Multimedia (ACM-MM'14)*. Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[43] Mountain Network. *Noardwand - Klimcentrum Leeuwarden*. May 2023. URL: https://mountain-network.nl/klimcentra/locaties/klimcentrum-leeuwarden/.

[44] VDiff. *How to belay: Top rope basics - learn to rock climb*. July 2023. URL: https://www.vdiffclimbing.com/basic-top-rope-belay/.

[45] Eduardo Fonseca et al. "Freesound Datasets: A Platform for the Creation of Open Audio Datasets". In: Oct. 2017.

[46] Gripped. *Indoor Weekly: Five tips for new route setters*. Apr. 2018. URL: https://gripped.com/profiles/indoor-weekly-five-tips-for-new-route-setters/.

[47] Gautham J Mysore. "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? A dataset, insights, and challenges". In: *IEEE Signal Processing Letters* 22.8 (2014), pp. 1006–1010.

[48] *BBC Sound Effects Library*. http://www.sound-ideas.com/sound-effects/bbc-sound-effects.html. URL is no longer available. 2015.

[49] Yi Hu and Philipos C Loizou. "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2007), pp. 229–238.

[50] Gabriel Mittag et al. "NISQA: A Deep CNN-Self-Attention model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets". In: *arXiv preprint arXiv:2104.09494* (2021).

[51] Antony W Rix et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 749–752.

[52] John HL Hansen and Bryan L Pellom. "An effective quality evaluation protocol for speech enhancement algorithms". In: *Fifth international conference on spoken language processing*. 1998.

[53] Cees H Taal et al. "A short-time objective intelligibility measure for time-frequency weighted noisy speech". In: *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2010, pp. 4214–4217.

[54] Ryandhimas E Zezario et al. "STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 482–486.

[55] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[56] Timo Gerkmann, Martin Krawczyk, and Robert Rehr. "Phase estimation in speech enhancement—unimportant, important, or impossible?" In: *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE. 2012, pp. 1–5.

A

Appendix I: Replicable Literature Review

Table A.1 shows how a large portion of the resources were found. Any sources not mentioned were either recommended by my external supervisor, or mentioned within one of the sources listed (or concern relevant non-academic sources, such as informal blogs about climbing). It should also be noted that search engines other than SmartCat and Google Scholar were used (such as ArXiv) but the results on Google Scholar generally proved to be the most relevant for my topic. ***Disclaimer:*** *sources found on Google Scholar and Google are often not peer-reviewed and might not be particularly trustworthy. However, every source mentioned below has been checked on this for academic validity.*

Search Engine	**Keywords**	**Source Name**	**Ranking**
SmartCat	"speech enhancement"	[18]	Top 2
SmartCat	"speech enhancement"	[12]	Top 2
Google Scholar	"speech enhancement" (2018 onwards)	[14]	Top 2
Google Scholar	"speech enhancement" (2018 onwards)	[11]	Top 2
Google Scholar	"self-supervised learning speech enhancement"	[6]	Top 2
Google Scholar	"self-supervised learning speech enhancement"	[15]	Top 2
Google Scholar	"survey speech denoising"	[16]	Top 1
Google Scholar	"speech enhancement time domain time-frequency domain"	[21]	Top 4
Google Scholar	"speech enhancement time frequency"	[10]	Top 1
Google Scholar	"speech denoising overview"	[19]	Top 1
Google Scholar	"unsupervised speech enhancement survey"	[13]	Top 1
Google Scholar	"speech enhancement phase"	[9]	Top 3
Google Scholar	"speech enhancement phase"	[27]	Top 3
Google Scholar	"speech enhancement time-domain survey"	[17]	Top 3
Google Scholar	"speech enhancement phase ambiguity"	[8]	Top 1
Google Scholar	"speech enhancement tf complex mapping"	[22]	Top 5
Google Scholar	"speech enhancement tf complex mapping"	[30]	Top 5
Google	"why is phase estimation difficult speech enhancement"	[56]	Top 1

Table A.1: Replicable Literature Search

B

APPENDIX II: TRAINING LOSS

This Appendix displays the training loss for the baseline model (**M1**). Throughout the training phase of each model experiment, the training loss over all epochs was monitored. The training loss provides an important indication of how well the model is learning from the training data, with a lower loss indicating improved learning proficiency. The results for **M1** can be seen in Figure B.1, with the training loss for the Clean Autoencoder (CAE) on the left, and the loss for the Mixture Autoencoder (MAE) on the right.

Even though the initial configuration of Wang et al.'s model [6] specifies 700 epochs for the CAE and 1500 epochs for the MAE, it is evident from the monitored training loss that this number is excessive. The loss for both autoencoders stabilizes before reaching 250 epochs and shows no further improvement. As the number of epochs can significantly increase the training time of the model, we find it important to briefly highlight these findings here.

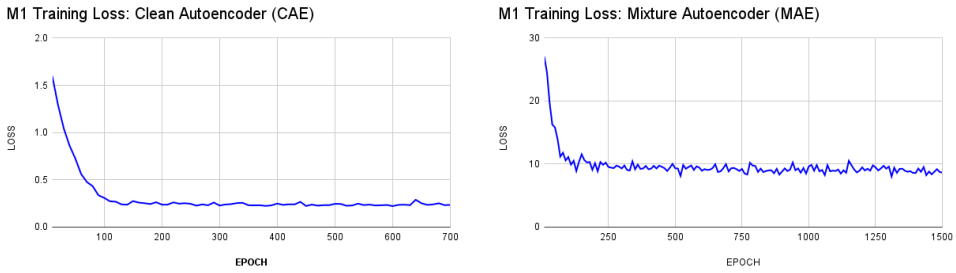


Figure B.1: Training loss CAE and MAE of **M1**