Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Henrik Hansen Stormyhr

A comparative analysis of MOS-LQO algorithms for perceptual transparency testing in audio steganography

Master's thesis in Information Security
Supervisor: Tjerand Silde
Co-supervisor: Bor de Kock & Emil August Hovd Olaisen
June 2025

NTNU
Norwegian University of
Science and Technology

Henrik Hansen Stormyhr

A comparative analysis of MOS-LQO algorithms for perceptual transparency testing in audio steganography

Master's thesis in Information Security
Supervisor: Tjerand Silde
Co-supervisor: Bor de Kock & Emil August Hovd Olaisen
June 2025

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

A comparative analysis of MOS-LQO algorithms for perceptual transparency testing in audio steganography

Henrik Hansen Stormyhr

10th of June 2025

Abstract

Steganography is the art of embedding a secret message into an inconspicuous container in order to conceal its existence. Unlike cryptography, steganography does not aim to obscure the contents of a message, but rather to hide the fact that a message exists at all. The field of audio steganography focuses on hiding such information in audio files, and a plethora of tools and methods have been created for this purpose. An important metric for evaluating these tools and methods is perceptual transparency. This metric says something about how difficult it is for a human observer to hear that a secret message has been embedded into an audio file. One way to measure this metric is by the use of Subjective Mean opinion score (MOS-LQS) tests and Objective Mean opinion score (MOS-LQO) algorithms that try to mimic the results of such tests. In this study the MOS-LQO algorithms PESQ, ViSQOL Speech and ViSQOL Audio are compared to the results of a subjective degradation mean opinion score (DMOS) test for evaluating our four chosen audio steganography tools and methods: Steghide, Hide4PGP, Low capacity GAN (GAN Low) and High capacity GAN (GAN High). Signal to noise ratio (SNR) is also measured in order to assess previous work's proposal of a 30 dB threshold value for human perception.

Our results show that ViSQOL Audio correlates the closest to DMOS across all of our audio samples, with PESQ also getting a very high correlation to DMOS, and ViSQOL Speech falling behind the two others by a significant amount. ViSQOL Audio also has by far the smallest absolute deviation from DMOS, deviating, in the worst case, approximately three times less on average than PESQ and ViSQOL Speech, which achieve similar deviations to each other. In addition to this, ViSQOL Audio shows a tendency to give moderately stricter evaluations than DMOS in these worst case scenarios, while PESQ and ViSQOL Speech show a tendency to evaluate far more lenient. Our results ultimately lead us to recommend the use of ViSQOL Audio for perceptual transparency testing within the field of audio steganography.

The identified SNR threshold value for human perception from previous work is also disputed, as audio samples shown to have perceptible degradations by our subjective DMOS test, simultaneously achieve SNR values far above the claimed threshold.

Lastly, a mathematical error in a paper proposing a tan map based audio steganography method is also discovered and documented.

Sammendrag

Steganografi går ut på å gjemme en hemmelig melding i et annet medium, helst uten å endre det på en måte som kan oppdages av uvedkommende. I motsetning til kryptografi er ikke målet med steganografi å skjule selve innholdet i informasjonen man gjemmer, men heller å skjule at informasjonen eksisterer i det hele tatt. Lyd-steganografi feltet fokuserer på å gjemme slik informasjon i lydfiler, og det har blitt utarbeidet utallige metoder for å oppnå dette målet. En viktig metrikk for å evaluere lyd-steganografimetoder er "perceptual transparency". Denne metrikken sier noe om hvor lett det er for et menneske å høre på en lydfil at den har blitt endret på. En måte å måle dette på er ved å bruke subjektive "mean opinion score" (MOS-LQS) tester og objektive "mean opinion score" (MOS-LQO) algoritmer som prøver å emulere de subjektive testene. I denne studien sammenliknes resultater fra MOS-LQO algoritmene PESQ, ViSQOL Speech og ViSQOL Audio med resultatene fra en subjektiv "degradation mean opinion score" (DMOS) test for å evaluere våre fire valgte lydsteganografi metoder; Steghide, Hide4PGP, lav kapasitet GAN (GAN Low) og høy kapasitet GAN (GAN High). "Signal to noise ratio" (SNR) blir også målt for å sjekke om tidligere arbeids forslag om en terskel på 30dB for menneskelig hørbarhet stemmer.

Resultatene våre viser at ViSQOL Audio korrelerer best med DMOS for alle lydklipp, og at PESQ også har en veldig høy korrelasjon til DMOS, ViSQOL Speech faller derimot bak de to andre med en betydelig lavere korrelasjon enn de to andre. ViSQOL Audio har også de klart laveste absolutte avviket fra DMOS, med et avvik som i verste fall er rundt tre ganger lavere enn PESQ og ViSQOL Speech sine avvik, som begge oppnår omtrent like avvik som hverandre. I tillegg til dette har ViSQOL Audio en tendens til å gi moderat strengere vurderinger enn DMOS, mens PESQ og ViSQOL Speech har en tendens til å gi mye mindre strenge vurderinger. På grunn av resultatene våre ender vi til slutt opp med å anbefale ViSQOL Audio for "perceptual transparency" testing i lyd-steganografi feltet.

SNR resultatene våre motsier også terskelen for menneskelig hørbarhet fra tidligere arbeid, da lydklipp med bevist stor nedsatt lydkvalitet fra den subjektive DMOS testen vår, oppnår SNR resultater langt over den foreslåtte terskelen.

Til slutt blir det også funnet og dokumentert en matematisk feil i en "tan map" basert lydsteganografi metode.

Acknowledgments

I would like to express my utmost gratitude to everyone that has helped make this thesis possible.

Firstly, I would like to give a big thanks to my supervisors Bor de Kock, Emil August Hovd Olaisen and Tjerand Silde for the great support they have given me throughout the project.

I would like to give a special thanks to Bor de Kock and Emil August Hovd Olaisen for attending weekly meetings with me during the entire project and pre-project. Their feedback and support has been invaluable. Both when it came to scoping down and defining the project, and while actually executing the plans from the pre-project and writing the thesis.

I would also like to give a thanks to all the participants who gracefully and patiently lent me their time by volunteering to join my Subjective DMOS test. This was a key part of the thesis, and the results and findings of my thesis could not have been discovered without their participation.

My close friends and family also deserve a big thank you for helping me recruit participants from different gender and age groups across their networks of acquaintances.

Contents

Figures

Tables

Chapter 1

Introduction

This chapter begins with introducing the Motivations behind this thesis. Then it describes the research problem and objectives, along with defining our two research questions to be answered. It then goes on to outline the specific contributions found by our study, before finally including a reflection on the thesis contributions towards sustainable development, in relation to the United Nations 17 goals for sustainable development [1].

1.1 Motivation

Audio steganography is the art of embedding a secret message inside an audio container [2]. It is customary to evaluate audio steganography in regards to three different metrics: perceptual transparency, hiding capacity and robustness [2]. In this thesis we will focus on different ways to evaluate the perceptual transparency of audio steganography methods.

An audio steganography method's perceptual transparency quantifies how much it alters the perceptible audio quality of its container file, in a way that can be heard by a human observer [3]. One way to measure this is by the use of a Subjective Mean Opinion Score (MOS) test [4], where participants typically rate the perceived audio quality on a scale from one to five [5]. Lower average ratings from the participants of this test would in turn translate to worse perceptual transparency, since a human can perceive that the audio steganography method has altered its container.

Another way to measure perceptual transparency is by the use of algorithms trying to emulate these types of subjective MOS tests. These are often referred to as MOS-LQO [6] algorithms, and the most commonly used one within the field of audio steganography appears to be PESQ [2, 7]. These algorithms are more convenient and less time consuming to use than arranging a Subjective MOS test, as you do not need to invite any participants. However, we have been unable to identify any previous work investigating what MOS-LQO algorithm is the most suited for evaluating the perceptual transparency of audio steganography

methods. This thesis aims to change this by putting a few different MOS-LQO algorithms to the test for this purpose.

Another metric that is often used to measure perceptual transparency within the field is Signal to Noise Ratio (SNR) [2, 3]. We have discovered a claimed SNR threshold for human perception in the literature [3, 8] that we also want to put to the test in this thesis.

Parts of this section was adapted from the IMT4205 Pre-project course report for this thesis, where this project was initially planned out [9].

1.2 Research Problem and Objectives

This thesis has a couple of objectives that it aims to investigate. The first objective is based on an observation made during the *IMT4205 - Research project planning* pre-project [9] for the Master's thesis. During this course it was discovered that the most commonly used MOS-LQO algorithm for evaluating the perceptual transparency of audio steganography methods appeared to be PESQ. This was discovered by surveying different literature such as this review [2] showing that PESQ was used in 12 out of 134 papers and PEAQ [10] in 3 out of 134. 29 out of 134 papers use either PESQ, PEAQ, Subjective MOS or a combination of the three, and 14 of the papers used MOS and neither of these two MOS-LQO algorithms [2]. No other MOS-LQO algorithms than PESQ and PEAQ was mentioned in the review. Google Scholar was also used to search for papers mentioning different algorithms identified during the pre-project [9] in relation to audio steganography; such as POLQA [11], ViSQOL [12] and AqUA [13], but no examples of other MOS-LQO used for evaluating perceptual transparency was found.

Several of these MOS-LQO algorithms are not free to use, and while a gray-zone in the PESQ license appears to allow for our specific academic use by allowing us to evaluate its performance [14], we had a hard time finding a license for PEAQ at all. We also tried to acquire academic licenses for POLQA and AqUA without luck. Google's open source MOS-LQO algorithm, ViSQOL, on the other hand, is free to use for anyone. The MOS-LQO algorithms PESQ, ViSQOL Speech and ViSQOL Audio will therefore be tested in this thesis.

This gray-zone license, combined with he fact that PESQ was released in 2001 [15], while ViSQOL was released in 2015 [12] makes us curious whether any of the two ViSQOL MOS-LQO algorithm modes (speech and audio) could be better suited than PESQ for evaluating the perceptual transparency of audio steganography methods. Which is also what we aim to investigate in this thesis.

The main goal of this thesis is to evaluate our chosen MOS-LQO algorithms by seeing how they compare to a subjective DMOS test. Subjective MOS tests are often seen as the "ground truth" for MOS-LQO algorithms [12], as they derives their results from real human perception [5]. The most commonly used subjective MOS test is ITU-T P.800 ACR procedure [5, 12], this procedure is typically referred to as just "MOS" [6]. In this study we are using the P.800 DCR procedure, typically referred to as "DMOS" [5], for reasons that are explained in Section 3.2. This pro-

cedure is based on the more commonly used ACR procedure, but is more sensitive to degradations in audio quality [5]. This paper about Korean synthesized speech shows a high correlation between the scores resulting from both procedures [16].

The thesis achieves its main objective of evaluating the chosen MOS-LQO algorithms by conducting a subjective DMOS test with 21 human participants following the P.800 DCR procedure [5]. In the DMOS test the participants rate the degradation in audio quality for audio samples containing information embedded by our chosen audio steganography methods, compared to a high quality reference, on a scale from 1-5. The average ratings from the DMOS test are then used to produce DMOS scores. The MOS-LQO algorithms are then applied to the same audio samples to give them MOS-LQO scores from 1-5, and the scores are compared by Pearson's correlation, mean absolute errors (MAE) and manual inspection.

Different statistical methods are also used to justify the exclusion of outliers from the DMOS test, and to see whether we can combine the DMOS scores from our male and female samples, or if we have to report them separately, with this last step of being recommended by the P.800 [5]. MOS-LQO algorithms that produce scores that both correlate closely with, and that are as close as possible to the DMOS scores, in terms of MAE, are likely the most desirable to use for evaluating the perceptual transparency of audio steganography methods. We elaborate on why we think this is the case in Section 3.2 of the thesis.

Signal to Noise ratio (SNR) is another common way to measure the perceptual transparency of audio steganography methods [3]. In fact, as much as 65 of the 134 papers reviewed by this paper [2] investigating the state of the art within the field in 2020 have included SNR as an evaluation metric. We think that SNR is likely more popular than MOS-LQO algorithms because of its ease of use and implementation, but the fact that it is not based on human perception [17] makes us question its suitability for evaluating perceptual transparency. While investigating this directly is not one of the goals of this thesis, some of our results did exaggerate our skepticism towards using SNR for this purpose, which we discuss further in Section 5.3. We therefore also propose this to be investigated further by future work.

During this project we also identified a claimed SNR threshold for human perception at 30 dB [3, 8]. This claim did not have any source or elaborating information to back it up, but had already made it into several papers [3, 8]. We therefore want to investigate the legitimacy of this threshold as a secondary objective of this thesis. This will be done by measuring the SNR of the same audio samples as the subjective DMOS test, and comparing the results of the two to see whether the threshold holds up or not, when compared to a metric based on real human perception.

For these objectives we have formulated the following research questions:

1. How do the different MOS-LQO algorithms; PESQ, ViSQOL Speech, and ViSQOL Audio compare to a subjective DMOS test, when it comes to evaluating the perceptual transparency of our chosen audio steganography methods, and which one appears to be the best suited?

2. How do the SNR scores compare to the subjective DMOS scores from our experiments, do the results support or oppose a threshold of 30 dB for human perception?

Our research questions are based on, but not exactly the same, as the research questions proposed in the pre-project report for this thesis made during the *IMT4205 - Research Project Planning* course at NTNU [9].

1.3 Contributions

The thesis produces several interesting results and contributions to the field. It is identified that ViSQOL [12] Audio appears to produce by far the closest results to our Subjective DMOS test, while PESQ [18] and ViSQOL Speech both produce results that differ significancy more from DMOS for our tested audio steganography methods than ViSQOL Audio. Suggesting that ViSQOL Audio should likely be chosen over the others when it comes to evaluating the perceptual transparency of audio steganography methods. ViSQOL Audio also consistently evaluates degradations in audio quality more strictly on average than our Subjective DMOS test, while ViSQOL Speech and PESQ consistently evaluate this less strictly. Since audio steganography methods are typically used to hide and secure secret information, we think that a more strict way of objectively evaluating the method's perceptual transparency is likely preferable to the opposite. So, this is another potential advantage of using ViSQOL Audio over the others. In addition to this, the observed less strict evaluations of PESQ and ViSQOL Speech are about three times further away from DMOS than ViSQOL Audio's worst case stricter evaluations, which we deem to be highly in favor of ViSQOL Audio. Combined with the apparent licensing issues of PESQ, our results strongly suggest that ViSQOL Audio should replace PESQ as the "go-to" MOS-LQO algorithm for perceptual transparency testing within audio steganography.

This finding is significant, as PESQ, which is identified as seemingly being the most used MOS-LQO algorithm within the field in [2], has a license that strictly speaking means it requires payment to use for purely evaluating audio steganography methods, even academically. ViSQOL on the other hand, is free and open source, and while ViSQOL Audio strongly outperforms PESQ in our study, we could still not find any papers using this MOS-LQO algorithm to evaluate the perceptual transparency of audio steganography methods. For the reasons mentioned above, we think that using ViSQOL Audio to evaluate the perceptual transparency of audio steganography methods rather than PESQ would likely be beneficial for the field.

In addition to this, the results from comparing our Subjective DMOS scores to our measured SNR values strongly suggest that a claimed SNR threshold of 30 dB for human perception discovered in the literature [3, 8] is incorrect. With audio samples proven by very low DMOS scores to be not only perceptible, but borderline annoying, receiving SNR scores far above 30 dB, where higher SNR

decibel scores typically translate to better perceptual transparency [3]. We therefore suggest future work exploring whether such a threshold makes sense, and if it does trying to identify one, as well as work exploring if SNR is really suitable for evaluating the perceptual transparency of audio steganography methods.

Lastly, a mathematical error is discovered in a paper [19] describing a Tan map based audio steganography method. This error was discovered by implementing the method (with help from Microsoft's CoPilot AI Chat-bot [20]) and seeing that we were not able to reproduce the results listed in the paper when using the same input for its functions, despite deterministic functions being a key feature of the method. Eventually, the specific error of missing a "+1" in an one of the functions, to offset the iterations of the two functions of the two dimensional TAN map used in the paper was discovered. This was noticed by finding a paper [21] describing a very similar method that the TAN map paper [19] was likely inspired by as they cited it in one of their previous papers using a sine map for audio steganography [22], that was cited in the tan map paper. This mistake was confirmed when correcting the issue in the implementation allowed us to produce most of the same results as the TAN map paper (one output value differed, but we strongly suspect this to be another spelling error in the tan map paper). This finding is included in the thesis as we deem that it could be useful for other people trying to troubleshoot why they are unable to reproduce the results of the TAN map paper.

1.4 Reflection on Sustainability

The NTNU resource on sustainability in computing describes sustainability as having three pillars; environmental impact, economic sustainability, and social responsibility [23]. In this section we will reflect on how the results of this thesis could affect the three pillars of sustainability, seen in relation to the 17 Goals for sustainable development set by the United Nations [1].

Better evaluation of audio steganography methods could make it easier for researchers to evaluate such methods, which could in turn save them time, and as we all know time is a valuable resource that is closely connected to the economic pillar of sustainability. Currently, the field does not seem to agree on a single subset of methods to be used for evaluating the perceptual transparency metric, work like this thesis could help bring the field together and strengthen collaboration by investigating what methods are the best to use for this purpose. This could in turn lead to research being able to be compared more easily without having to run extra tests, potentially saving researchers even more time.

Our results also show that the best performing MOS-LQO algorithm of the ones tested appears to evaluate perceptual transparency either very similarly, or moderately stricter than a subjective DMOS test with 19 human participants. With the stricter results happening for audio files that were already very perceptible according to the DMOS test. These subjective DMOS tests can be costly and time consuming to conduct, so research like this thesis helping to find objective al-

gorithms that are free and quick to run could further contribute to the economic pillar of sustainability. In addition to this, our findings suggest that a free and open source alternative MOS-LQO algorithm, that we could not find any existing audio steganography papers using, appears to significantly outperform a paid algorithm that is often used in the field, contributing directly to the economic pillar of sustainability.

In addition to this, we can also find some more indirect social and environmental benefits of this work. The work done in this thesis could help actors trying to protect highly sensitive information by the use of audio steganography, evaluate the perceptual transparency of methods they are considering for this purpose. Depending on their use of this audio steganography method, one could picture several scenarios where this could benefit both social and environmental sustainability. One type of actor that could potentially benefit from the added security provided by secret communication through audio steganography could be political dissidents in oppressive regimes. The work done in this thesis could help such actors pick a more secure method for communication by helping them select an audio steganography method with good perceptual transparency. Helping such actors communicate secretly without being discovered is a clear social benefit that could be helped by our work. Another example could be if an oil company wanted to use some type of real time VOIP based audio steganography communication system for critical and sensitive communication. The work done in this thesis could help this company pick the right method in terms of having good perceptual transparency, and potentially indirectly help avoid environmental disasters caused by attacks from different types of threat actors.

The things discussed here contribute to goal 9 and 16 of the United Nations 17 goals for Sustainable Development, by contributing to the following targets: "9.1 - Develop quality, reliable, sustainable and resilient infrastructure, including regional and transborder infrastructure, to support economic development and human well-being, with a focus on affordable and equitable access for all" [24], "9.5 - Enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending" [24], and "16.10 - Ensure public access to information and protect fundamental freedoms, in accordance with national legislation and international agreements" [25].

All in all, this thesis could have several direct and indirect benefits to all the three pillars of sustainability. While some of these types of benefits could be shared by much of the work done in the field of information security, some of them are also extra relevant to this thesis's work in particular. The thesis also contributes to several of the United Nations 17 goals for sustainable development.

Chapter 2

Background

This chapter presents the necessary background information on the concepts used and discussed in this thesis. It starts by explaining the concept of steganography, before going on to further clarify what audio steganography is. It then introduces the typical ways of evaluating audio steganography methods, stating that we will focus on the perceptual transparency metric in this thesis. The audio steganography tools and methods used in the thesis are then introduced and explained in detail. Then, the concepts of subjective and objective mean opinion score (MOS) testing are explained, along with different ways of doing this. This includes the explanation of the absolute category rating and degradation category rating, as well as introducing and providing background information the different MOS-LQO algorithms used in this study, and explaining why some MOS-LQO algorithms had to be excluded. Finally, the concept of signal to noise ratio (SNR) is explained.

2.1 Steganography

Steganography can be found in many varieties, but it typically involves disguising a secret message by embedding it inside an inconspicuous container of some sort [26]. In contrast to cryptography, steganography typically does not hide the contents of the secret message, but rather the fact that a message has been sent at all [2]. Because of these complimentary qualities, steganography is often combined with cryptography when deployed to achieve stronger security by hiding both the contents of a message and the fact that the message has been sent [2]. You cannot try to crack a message that you don't know has been sent, and it might also be harder to extract a steganography message from its container if the content appears to be random.

The earliest recorded instance of steganography is from Greece in year 440 BC, where the Greek ruler Histiaeus sent a message to his vassal by marking a message on one of his loyal men's shaved scalp [26]. Once the hair regrew the message was hidden and could be retrieved by shaving the head again. Steganography has historically often been used in wars, with some notable examples being the use of invisible ink by prisoners in war camps in WW2 [26], and Jermiah Denton blinking

in morse code after being captured and forced to participate in a propaganda interview to convey a message of torture being used in prison camps during the Vietnam war [27]. Another somewhat well known modern example could be the yellow dot-matrix code used in some modern laser printers to encode timestamps, model and serial numbers onto printed documents in a way that is invisible to the naked eye [28]. Recently, there has also been much discussion about using steganography to watermark AI-generated content [29].

In modern electronic steganography the host medium is often referred to as the "cover file" and the altered cover file containing the secret embedded message as the "stego file" [2]. These terms will also be used in this thesis as we are applying modern electronic steganography methods that use audio files as their cover medium. We will also sometimes refer to steganography methods as "stego methods".

2.2 Audio Steganography

Audio steganography is a form of steganography that uses audio as its cover medium and that can be described as the art of embedding a secret message into an audio container [2]. In this thesis, we are focusing on audio steganography methods utilizing audio files specifically, which appear to be most of the modern audio steganography methods [2]. However, audio steganography methods utilizing non-file cover mediums such as continuous VOIP (Voice over IP) transmission also exist [2].

2.2.1 Perceptual Transparency, Robustness and Hiding Capacity

Audio steganography methods typically balance a trade off between three different qualities: Perceptual transparency, hiding capacity and robustness [2]. These are therefore also typically the metrics that are used to evaluate the performance of audio steganography methods upon their creation or later evaluation [2].

Hiding Capacity has perhaps the most obvious meaning of the three and refers to how much information that can be embedded into the cover file by an audio steganography method [3]. While robustness refers to how well the stego file containing some secret information holds up to different types of attacks or modifications [3]. Some examples of "attacks" against audio steganography stego files could be compression, noise addition and audio filters [3].

The metric we are the most interested in this thesis is perceptual transparency. This metric refers to the similarity between the cover file and the stego file [2]. It can also be described as how easy it is for a human observer to hear that a hidden message has been embedded into the stego file [3]. The fact that the word "perceptual" is used here makes us like this last definition of the metric better. However, this review of 134 audio steganography methods shows that SNR is a more popular way to measure perceptual transparency than subjective MOS tests and objective MOS-LQO algorithms. With 65 of the papers using SNR to measure

perceptual transparency, which is not based on human perception, but rather the difference in noise between the cover and stego files [17]. 29 of the papers in the review [2] are using either MOS [5], PESQ [7] or PEAQ [10] to evaluate perceptual transparency, which are all methods that are based on or meant to emulate human perception.

2.3 Audio Steganography Tools and Methods Used

This thesis implements two audio steganography methods from 2019 and applies two older audio steganography tools from the early 2000's. These tools and methods were chosen during the pre-project [9] to be the same tools and methods used in Reyer's bachelor thesis [3] which evaluates the hiding capacity, robustness and perceptual transparency of these methods in a purely objective manner. This selection was in part done because Reyer's thesis greatly inspired the theme for this thesis by suggesting the exploration of MOS-LQO algorithms like PESQ for evaluating the perceptual transparency of audio steganography methods. Reyer's thesis focuses on evaluating the steganography methods themselves, and uses signal to noise ratio (SNR) instead of perceptual models like MOS-LQO algorithms to evaluate their perceptual transparency.

In contrast to Reyer's thesis [3], our thesis focuses on evaluating different MOS-LQO algorithms, which are models based on human perception. The stenography methods used in Reyer's thesis are chosen because they appear to have done a thorough search when selecting these methods, stating that it was not easy to find recent audio steganography methods that were actually able to be implemented easily from their published papers. We also agreed with their logic of choosing two recent audio steganography methods from papers published after 2016, and two older well known audio steganography tools that have commonly been included in research. However, the methods are not applied exactly like Reyers and one method was excluded from the main experiments for reasons that will be explained in Section 3.5.4 and 4.5. The methods are described in the Subsections 2.3.1, 2.3.2, 2.3.3, and 2.3.4, following below.

2.3.1 Steghide

The first steganography tool used in this thesis is Steghide 0.5.1, created by Stefan Hetzl and released in 2003 [30]. Steghide's audio steganography feature works by first splitting the selected cover audio file into smaller samples, the positions of the samples that will have the cover file embedded into their least significant bits (LSB), i.e. the last bit of the sample, are chosen by a pseudo-random number generator with a passphrase as the input [31]. The positions already containing correct bit-values by chance are filtered out and a graph-theoretic pair matching algorithm is used to find pairs of samples and parts of the cover data to be embedded that can be swapped without interfering too much with the first order statistics of the stego file [30]. In the end the data that did not have any pairs

identified by this algorithm is overwritten on to remaining pair-less samples [31]. This method is used to preserve first-order-statistics in order to make it more difficult to detect that a secret message has been embedded into the stego file [31]. More technical details about our application of Steghide can be found in Section 3.5.1.

2.3.2 Hide4PGP

The second steganography tool used in this thesis is Hide4PGP 2.0, this tool was created by Heinz Repp in 2000 [32]. The basics of how the audio steganography part of the Hide4PGP steganography tool works is explained in the manual file included with the download [32]. The manual explains that Hide4PGP is meant to be used with PGP and that the cover files needs to be encrypted with PGP by the user manually before using Hide4PGP to embed them into the stego file. The manual also explains that Hide4PGP splits up audio files into their samples and that it can change up to 1 bit per sample in VOC files and 8-bit WAV files and up to 4 bits per sample in 12 and 16-bit WAV files, the manual states that these limits were put in place to ensure that the average listener would not be able to hear that an embedding of information had taken place in the stego file. It also elaborates that the embedded data is spread evenly across the stego file. The manual further explains that Hide4PGP only works on "real" data and doesn't modify file headers etc. so that lossless file conversions won't ruin the embedded data.

It was difficult to find resources explaining exactly how Hide4PGP works, but according to [3] Hide4PGP uses a variation of least significant bit (LSB) substitution [33] when used on audio files. [34] also states that Hide4PGP uses a variation of LSB substitution when used on image files. Combining this information with the previous information from the manual we find it likely that Hide4PGP works by evenly substituting the least significant bit of each audio sample when embedding information in VOC and 8-bit WAV files, and the four least significant bits when embedding in 12 and 16-bit WAV files. [3] achieves almost exactly 25% hiding capacity both while embedding in 16-bit WAV files from the GZTAN [35] and TIMIT [36], and we also achieve similar hiding capacities to this when applying the tool on our 16-bit WAV files, which coincides with the Hide4PGP manual [32] stating that four bits per sample are used for embedding information in 16-bit WAV files.

2.3.3 GAN Based Audio Steganography

In this thesis we refer to "GAN Based audio steganography" or "GAN Low" and "GAN High" as the audio part of the combination of the two steganography methods first proposed the papers [37] and [38], and then combined, extended and implemented in [39]. GAN Low and GAN High refers to this same method applied at low and high embedding capacities, this is further explained in Section 3.5.3. This method uses three generative adversarial networks (GANs) [40] to embed information into images and audio files [39]. The method uses an encoder, decoder

and critic approach with all of these components being their own GAN. The encoder encodes data into a file, the decoder decodes them, and the critic attempts to evaluate how easily the presence of a secret message can be detected in the generated stego file in order to improve its quality [39]. For audio steganography, the cover audio files are turned into their respective spectrograms before being passed through the GAN for embedding the secret messages, after the embedding has taken place they are turned back into audio files [3]. The spectrograms generated for one of our audio files can be seen in Figure 2.1.

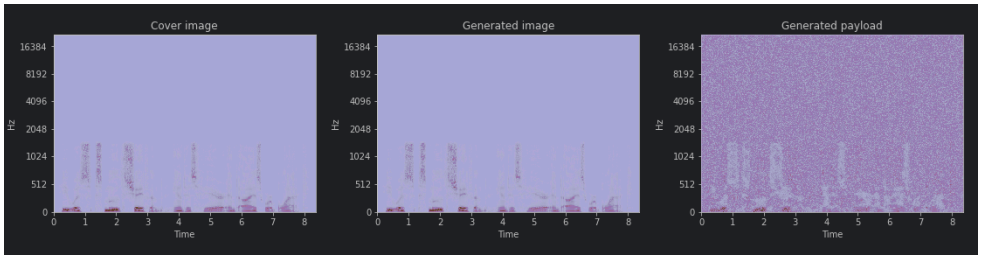


Figure 2.1: The spectrograms generated by the GAN method for one of our audio files.

2.3.4 TAN Based Audio Steganography

When referring to "TAN based audio steganography" or the "TAN method" in this thesis, we are referring to the audio steganography method proposed in [19]. This method first uses basic AES-128 [41] encryption before applying a chaotic 2D TAN map system in order to decide what order of 16-bit WAV file audio sample's least significant bits (LSB's) to use for embedding the hidden message. This system uses two mathematical functions to make up the 2 dimensional TAN map that will generate chaotic looking outputs from their inputs, applying a similar idea to the pseudo random number generators often used in encryption. The chaotic behavior of the 2D TAN map is displayed in Figure 2.2 by plotting the values returned by its functions. The output of each function is used to decide the order samples whose LSB's are to be used for embedding the hidden message, with the output of one function deciding the positions of samples in the left audio channel of a stereo file and the other deciding positions for the right channel [19]. The reasoning behind using a 2D TAN map for this system is that they are deterministic, but generate widely different outputs for even a small change of the input value. In practice this means that the starting input given to the functions can be used as a key for being able to extract the hidden message from the stego file, providing extra security to the method [19].

The paper proposing the TAN method [19] has a mathematical error which was not discovered in time to include the method in the main experiments of this thesis. However, we consider the discovered error and how to solve it as a meaningful result and therefore include and document it in this thesis. More inform-

ation about this mathematical error and the implementation of the TAN method can be found in Sections 3.5.4 and 4.5.

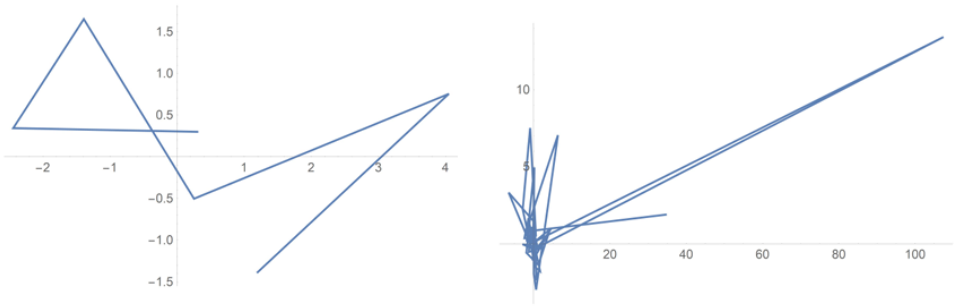


Figure 2.2: The chaotic behavior of the 2D TAN map used in [19] after fixing the mathematical error, in the first 5 iterations (left plot) and the first 40 iterations (right plot), adapted from [19].

2.4 Mean Opinion Score Testing

A Mean opinion score (MOS) test as described in the *ITU Telecommunication Standardization Sector's (ITU-T) P.800: Methods for subjective determination of transmission quality recommendation (P.800)* was originally just a subjective way of measuring the perceived audio quality of a telecommunication system [5]. The idea of a MOS test as described in the P.800 recommendation is to rate the perceived audio quality of a voice transmission system [5]. In all objective and subjective MOS procedures and algorithms relevant to this thesis, this audio quality is ranked on on a scale from 1-5. However, the P.800 also contains other scales such as the comparison category rating (CCR) scale ranging from 3 to -3 [5], but these are not relevant for this thesis as we have not found objective MOS algorithms using other scales than 1-5.

The P.800 recommendation proposes procedures for five different styles of Subjective MOS testing, where some involve subjects engaging in conversation, while others are pure listening tests [5]. Since the release of the P.800 recommendations many objective algorithms trying to emulate these types of subjective MOS tests have emerged, such as PESQ [18], POLQA [11], ViSQOL [12] and AqUA [13]. In addition to this, MOS testing has also started to be used in the field of video quality [6]. Likely because of this broadness, the ITU-T proposes some new, more specific, terminology to be used for the different types of MOS testing in the *P.800.1: Mean opinion score (MOS) terminology recommendation* [6]. The P.800.1 recommendation proposes different abbreviations to be used for different types of MOS testing like audio and video MOS tests, conversation and listening based MOS tests and subjective and objective MOS tests [6].

Some abbreviations from the P.800.1 recommendations [6] that are relevant

for this thesis are the ones for listening based subjective and objective MOS tests, which are defined as "MOS-LQS" (MOS subjective listening quality) and "MOS-LQO" (MOS objective listening quality) respectively. The MOS-LQO abbreviation from the P.800.1 recommendation refers to an objective MOS test performed by an algorithm, while MOS-LQS refers to a subjective MOS test that has been done using the absolute category rating (ACR) scale from the P.800 recommendation [5], this procedure will be explained further in Section 2.5.1. In addition to this, the P.800 recommendation's "DMOS" (degradation MOS) abbreviation, referring to a subjective MOS test conducted in accordance to the Degradation Category Rating procedure (DCR), is also highly relevant to this thesis [5]. The DCR procedure will be explained further in Section 2.5.2.

2.5 Subjective MOS testing

The ITU-T P.800 recommendation describes five different procedures for conducting subjective MOS tests [5]. Of these five, two of the procedures are particularly relevant for this thesis. These are the Absolute Category Rating (ACR) and Degradation Category Rating (DCR) respectively, these procedures follow the same core principles, but deviate in a few meaningful ways. Both of the procedures focus on rating the quality of audio from 1-5, but the formulations of the scales presented to the participants to rate by varies between the two [5]. Despite their differences, studies have previously shown high correlation between the two [16]. The ACR and DCR procedures will be explained in further detail in the Sections 2.5.1 and 2.5.2 following below. The selection of one of these procedures for our subjective MOS test will be done in Section 3.2, so they will also be discussed even further there.

2.5.1 Absolute Category Rating

According to [12], the Absolute Category Rating (ACR) procedure appears to be the most commonly used P.800 [5] procedure for measuring mean opinion scores (MOS) subjectively. This also appears to be the case for the field of audio steganography when looking at this review of the state of the art within the field from 2020 [2]. As explained in Section 2.4 on Subjective MOS testing, the ACR procedure ranks audio quality on a scale from 1-5. The ACR procedure contains three of these scales, where the "Listening-quality scale" appears to be the most commonly used [12], as displayed in the far right of Table 3.1, we will therefore focus only on this ACR scale in this thesis.

The scales of the different MOS procedures contain formulations given to the participants to instruct them on how to rate the perceived audio quality [5]. The scores produced by conducting a subjective MOS test using the Listening-quality scale with the ACR procedure are often simply referred to as MOS scores, as this is the terminology shown in the P.800 recommendation [5]. However, the P.800.1 recommendation [6], updating the recommended MOS terminology, refers to MOS

tests done using any of the ACR scales as MOS-LQS, which stands for "*mean opinion score listening quality subjective*" [6]. This is done to differentiate the subjective MOS tests from the objective MOS algorithms emulating these tests, which are defined as MOS-LQO or mean "*opinion score listening quality objective*" [6] in the P.800.1. However, at least within the field of steganography, it seems more common to refer to scores from ACR as MOS and using the names of the specific MOS-LQO algorithm used to describe their respective scores [2]. Nevertheless, we have seen most of these terminologies being used within the field [2] and we will be using all of them in this thesis as we see fit.

The P.800 ACR procedure contains many recommendations to be followed which will be explained in further detail in Sections 3.4 and 2.5. For now, the most important ones to showcase the difference between the scales used in ACR and DCR, and the fact that ACR does not use high quality reference audio samples, instead the participants rate the overall audio quality of the degraded audio samples only [5].

2.5.2 Degradation Category Rating

The P.800 degradation category rating (DCR) procedure is based on the ACR procedure, but differs in a few meaningful ways [5]. For DCR, high quality reference audio samples are used and the participants listen to these before listening to each of degraded audio samples. They then rate the degradation of audio quality from the high quality reference on a scale from 1-5, based on the DCR scale formulations shown in Table 3.1. The scores generated from a MOS test following the DCR procedure are usually referred to as DMOS scores, which stands for degradation mean opinion scores, we will be using this abbreviation in this thesis, both to describe a test done with the DCR procedure (DMOS test), and to describe the scores generated (DMOS scores). According to the P.800 recommendation, the DCR procedure has been shown to be more sensitive than ACR.

2.6 Objective MOS Testing

Objective MOS testing typically refers to models or algorithms used to emulate the results of Subjective MOS tests [12, 15]. Typically, these algorithms try to emulate the scores from a ITU-T P.800 [5] MOS test following the ACR procedure. According to the ITU-T P.800.1 recommendation, the scores generated from these algorithms can be abbreviated as MOS-LQO. We will therefore sometimes refer to these algorithms as MOS-LQO algorithms in this thesis, as it gives us an easy way to refer to all of them at once.

2.6.1 PESQ

PESQ is an older MOS-LQO algorithm that was first proposed in 2001 in the ITU-T P.862 recommendation [18] as the ITU-T's new recommended way of evaluating

speech quality objectively [15]. It has since been replaced by POLQA in 2011 [42] in the P.863 recommendation [11]. PESQ is mostly based on subjective P.800 [5] ACR MOS experiments [43], and originally only worked for narrow-band audio in the P.862 recommendation [18]. In the beginning, PESQ also provided scores on a range from -0.5 to 4.5 that were not directly comparable to ACR MOS scores [44]. However, this was later changed with a mapping function, mapping the raw PESQ scores into the ACR MOS range from 1-5 in the ITU-T's P.862.1 recommendation in 2003 [44]. When the wide-band version of PESQ was introduced in the P.862.2 recommendation [45] in 2007, this mapping function was included into the PESQ wide-band code. This means that PESQ measurements taken using the wide-band version of PESQ produces scores from 1-5 that are directly comparable to subjective ACR MOS scores.

When we refer to PESQ in this thesis we are referring to wide-band PESQ reimplemented to be used with a python wrapper in this GitHub repository [46]. This is a popular way to use PESQ now a days and big actors like NVIDIA and Facebook research can be seen using this repository [46]. More about our specific PESQ implementation can be seen in Section 3.7.2.

PESQ is first and foremost a paid piece of software, and a license needs to be followed in order to use it [14]. Even for academic use, most use cases call for obtaining a paid license [14]. However, there are some exemptions to this, such as being able to evaluate the performance of the algorithm's intended function [14]. Since our main focus in this thesis when it comes to PESQ is to evaluate how well it performs compared to other MOS-LQO algorithms, we have assessed that we are within this exemption. However, we are not convinced that papers simply using PESQ to evaluate their own proposed audio steganography methods are strictly speaking within their rights to do so without a paid license.

2.6.2 ViSQOL

Unlike PESQ, ViSQOL is a free and open source MOS-LQO algorithm developed by Google [47]. Like PESQ, ViSQOL also appears to be mostly based on subjective P.800 [5] ACR MOS experiments during its development, judging by the paper proposing it [12] only referring to the ACR procedure, calling it the most commonly used procedure, and doing an ACR MOS test as their only subjective comparison. ViSQOL is made as a free alternative to PESQ and POLQA, and like them it scores audio quality on a range from 1-5 [12].

ViSQOL has two modes; speech and audio, the original proposal only contained ViSQOL Speech [12], while the audio mode was introduced with ViSQOL V3 in a later paper [48] from 2020. We will refer to these modes as two separate MOS-LQO algorithms called ViSQOL Speech and ViSQOL Audio in this thesis. When referring to ViSQOL in this thesis we are referring to version 3.3.3 which is the latest released version on the ViSQOL GitHub as of the writing of this thesis [47], this is also the version we use in our testing. More information about the used ViSQOL implementation used in this thesis can be found in Section 3.7.3.

ViSQOL Audio (v. 3.3.3) technically does not follow the P.800 ACR rating scale to a tea, as it caps out at a MOS-LQO score of about 4.75. However, we do not think this is a problem as even our extremely hard to hear steganography methods, Steghide and Hide4PGP, did not achieve PESQ, ViSQOL Speech or ViSQOL Audio scores of higher than about 4.73 for any samples, with PESQ and ViSQOL Speech consistently ranking these samples lower than ViSQOL Audio. This combined with the fact that ViSQOL Audio correlated the closest to DMOS and got scores with the least absolute deviation from DMOS on average makes us think the 4.75 score cap is not a problem in practice. In addition to this, the average DMOS score of all null pair's, meaning we show the participants two of the same high quality reference samples then ask them to rate the degraded difference, is about 4.79 in our subjective test, further suggesting that the 4.75 cap for ViSQOL Audio is likely not a problem, as it is very close to the average DMOS score of a perfect sample.

2.6.3 Excluded MOS-LQO Algorithms

During the pre-project [9] we also identified POLQA [42] and AqUA [13] as potential MOS-LQO algorithms to include in our comparison. The POLQA Coalition's website states that OPTICOM [49] might be able to provide an academic license for POLQA [50]. However, when contacting them during the pre-project, we were told that they could not provide us with this for this thesis. We also contacted Sevana Öu [51], the creators of AqUA to ask for an academic license which informed us that they no longer offered academic licenses. OTPICOM informed us that an academic license for POLQA could potentially be acquired from one of its many vendors, but we decided that contacting 100+ vendors would be a poor use of our limited time for producing this thesis. Especially since ease of use and acquisition is likely also beneficial if we want to find a new standard MOS-LQO algorithm for perceptual transparency testing within the field of audio steganography.

PEAQ was also identified as a relevant MOS-LQO algorithm to include, as it is sometimes used for audio steganography perceptual transparency testing, although it appears to be quite a bit less commonly used than PESQ [2]. However, this algorithm was excluded as we were not able to find any licensing information for it other than it being a paid piece of software [52].

2.7 Signal to Noise Ratio

SNR appears to be the most commonly used way to measure perceptual transparency in audio steganography according to this review [2] of the state of the art within the field from 2020. We think that this method is likely so popular because it is quite easy to implement and not protected by any licenses. However, unlike MOS-LQO algorithms, it is not based on human perception [17]. We discuss this as a potential weakness of SNR when it comes to measuring the perceptual transparency of audio steganography methods in Section 5.3.

The basic definition of signal to noise ratio (SNR) is that it is a ratio between the power of a signal and the power of noise [17]. SNR is usually measured in decibel (dB) and its general definition can be expressed with this formula [17]:

$$SNR_{dB} = 10log_{10}\frac{P_{signal}}{P_{noise}}$$

In practice SNR can be measured in several ways, but for evaluating audio steganography a commonly used method [17] of looking at the difference between the cover and stego files and assuming the entire difference is noise, is likely the easiest to implement and use. This method also works particularly well for evaluating audio steganography methods since you usually have a clean cover file to use as your reference without added noise from the audio steganography method. By measuring SNR with this method you are therefore isolating the measurement to only measure noise added by the audio steganography method, which is likely what you would want from such a measurement. More information about our specific SNR implementation can be found in Section 3.8.

Chapter 3

Method

This chapter explains the methodology of the necessary activities done in order to answer our research questions. It also discusses and justifies the selection of these activities. The chapter starts by explaining our idea of using subjective mean opinion score (MOS-LQS) tests as a benchmark for our objective mean opinion score (MOS-LQO) algorithms. It then goes on to discuss considerations for choosing the right MOS-LQS procedure from the ITU-T P.800 [5] recommendations, concluding that the degradation mean opinion score (DMOS) procedure is the most suited for our purposes. Recruitment and practical considerations for the subjective DMOS test are then discussed, before explaining our method for selecting a dataset, and audio files to be used as the cover files for our audio steganography methods. The application and implementation of each of our chosen audio steganography methods is then explained in detail, followed by the DMOS experiment design. Then, the methods for applying and implementing our chosen MOS-LQO algorithms are described, as well as our method for implementing and measuring the signal to noise ratio of our audio samples. Lastly, the methods for calculating DMOS scores and analyzing and comparing our various results are presented.

3.1 Using MOS-LQS as a Benchmark for our MOS-LQO Algorithms

To find out what MOS-LQO algorithm is the most suited for evaluating the perceptual transparency of audio steganography methods we are planning to first conduct a subjective MOS test to use as baseline for comparison of the algorithms. As stated in the background section, MOS-LQO algorithms are typically made to emulate subjective MOS tests [12, 13, 15, 53], and the paper proposing ViSQOL states that subjective testing with human participants are to be considered the "ground truth" for these types of algorithms [12]. We therefore think that comparing performing a subjective MOS test that we can compare MOS-LQO algorithms to is a good place to start, when it comes to answering our first research question, that can be found in Section 1.2. This section describes the steps and considerations

taken to conduct the subjective MOS test.

3.2 Choosing the Right ITU-T P.800 MOS-LQS Procedure

As stated in Section 2.5, the ITU-T P.800 recommendations contains recommendations on how to design different types of subjective MOS tests [5]. However, there are still decisions and considerations that need to be taken while following the P.800, as it leaves many things up to the people designing the experiment. Looking at additional literature can also be helpful, as well as carefully considering the options best suited for our specific situation.

For instance, the P.800 differs between "Listening-opinion tests", which are MOS tests where the test subjects are purely listening to some audio samples and "Conversation-opinion tests", for example, which is another type of MOS test described in the P.800, as well as also describing several other types of MOS tests [5]. The subjective MOS test done in this thesis is a "Listening-opinion test" as described by the P.800. Even for this specific type of MOS-LQS test, several procedures are mentioned by the P.800. We deem the "Absolute Category Rating" (ACR) and "Degradation Category Rating" (DCR) procedures from the P.800 to be the most relevant for our purposes, as these both result in scores from 1-5, just like most MOS-LQO algorithms that we have been able to identify [12, 13, 15, 53]. Many MOS-LQO algorithms also appear to be based largely on ACR tests [12, 13, 15, 53], and DCR is based on ACR [5]. The difference in the purpose behind these procedures is highlighted by the language used to describe their rating scales from one to five [5], the formulations of these rating scales can be seen in Table 3.1.

It should also be mentioned that the ACR method includes two other rating scales, one for listening effort and one for loudness preference, the one displayed here is known as the "*Listening-quality scale*" [5] and is the one we assessed to be relevant to our study. When referring to the ACR method in this thesis we are referring to using the ACR method with the Listening-quality scale exclusively, as we found that this appeared to be the one mentioned in most of the MOS-LQO algorithm's we were able to identify papers and resources [12, 13, 15, 53].

Score	DCR	ACR
5	Degradation is inaudible.	Excellent
4	Degradation is audible but not annoying.	Good
3	Degradation is slightly annoying.	Fair
2	Degradation is annoying.	Poor
1	Degradation is very annoying.	Bad

Table 3.1: The different formulations used for rating audio quality in the P.800's DCR and ACR methods, adapted from [5].

The ACR procedure as described in the P.800 aims to assess the general speech or audio quality of some audio samples [5]. Because of this inherent goal, no ref-

erence audio is used in ACR MOS testing [5]. The goal behind the DCR procedure on the other hand, is to evaluate how much the quality has degraded from a high quality reference [5]. The P.800 states that this is likely one of the reasons that the DCR procedure has shown to be more sensitive than the ACR procedure, meaning that it should be better at detecting small changes in audio or speech quality [5]. In this thesis we are more interested in audio quality than speech quality, as we think this is more relevant to the degradations typically caused by audio steganography methods.

We assess that the increased sensitivity of the DCR procedure makes it the most suited for our purposes. One of the most important characteristics of audio steganography methods is that they should be difficult to detect. The ACR procedure's lower sensitivity has us concerned about it possibly not being sensitive enough to answer our research question. This thesis will use Subjective DMOS scores as a baseline to compare some chosen Objective MOS (MOS-LQO) algorithms. We think that it would be more difficult to draw sensible conclusions from this comparison if the chosen MOS-LQS procedure is not sensitive enough to pick up degradations in audio quality for audio steganography methods where they may be difficult, but still potentially possible, to hear.

However, PESQ which is one of the MOS-LQO algorithms that we plan to use was mostly compared to subjective MOS tests using the ACR procedure during it's development [43], and the paper proposing ViSQOL [12], which is another MOS-LQO algorithm we plan to use, also only refers to the ACR procedure, as well as using it in one of its experiments comparing ViSQOL to PESQ, POLQA and AN ACR MOS-LQS test. In addition to this, the P.800 states that the MOS notation is reserved for the ACR procedure using the listening-quality scale, while results from tests using the DCR procedure shall be labeled as "DMOS" scores [5]. As both wide-band PESQ and ViSQOL report their results with the MOS notation, it is likely that they are trying to emulate the ACR procedure. Despite this, we will use the DCR procedure for our MOS test because of the sensitivity concerns mentioned earlier. One paper [16] comparing DCR DMOS, ACR MOS and PESQ for evaluating Korean synthesized speech also found that DCR DMOS correlated closer to PESQ than ACR MOS, as well as ACR and DCR correlating closely to each other. The DeepL AI PDF translator [54] was used to translate this Korean article, but the abstract is in English and contains the information referred to here.

We argue that the DCR DMOS results will likely be more useful than the ACR methods MOS scores to find out what MOS-LQO algorithm appears to be best suited for evaluating the perceptual transparency of audio steganography methods. The same scale of one to five is used for both procedures and the P.800 states that DCR is based on ACR [5], DMOS should therefore simply be a more sensitive version of the ACR MOS score. When combining this with the results from the Korean study [16] mentioned earlier showing high correlation between MOS, DMOS and PESQ we conclude that comparing DMOS and MOS-LQO scores should not be any more or less problematic than comparing them to ACR MOS. Since there are also other to measure perceptual transparency other than MOS, such as

the SNR (signal to noise ratio) method used by Reyers [3], and the DCR DMOS procedure still uses human perception to measure degradation in audio quality, we find it just as suitable to measure the perceptual transparency as ACR MOS. We would also argue that it is desirable for a MOS-LQO algorithm to get as close results to DMOS as possible when evaluating the perceptual transparency of audio steganography methods, as a more sensitive algorithm could be useful when comparing audio steganography methods that get increasingly difficult to detect by the human ear. In addition to this, the added sensitivity of the P.800 DCR procedure is also deemed to be desirable by us when evaluating technology that is meant to secure sensitive information.

The paper proposing ViSQOL states that Subjective MOS tests are the ground truth for MOS-LQO algorithms [12], and as stated previously PESQ was also developed largely with ACR MOS scores in mind [43]. We therefore find it likely that our planned methodology of comparing different MOS-LQO algorithms to the Subjective DMOS scores of different audio steganography methods is an effective way to measure their suitability for measuring perceptual transparency within the field. This is further supported by the fact that the ViSQOL paper mentioned at the start of this paragraph also does something similar to our methodology by comparing the results from a subjective MOS test to results from the ViSQOL and PESQ MOS-LQO algorithms, on audio modified with degradations that are typical for VOIP systems, in order to see what algorithm performs the best in different scenarios.

While the DCR procedure is ultimately chosen for this study, it is less defined in the P.800 than the ACR procedure [5]. Because of this, we will also try to follow the ACR procedure when something is not defined in the DCR procedure, but is defined in the ACR procedure. We think it makes more sense to do this rather than making up our own choices on the spot, especially since the DCR procedure is based on the ACR procedure, according to the P.800 [5]. In addition to this, it is not always possible for us to completely follow either procedure for practical reasons, in these cases we try to come up with solutions that are inspired by one of the procedures, but that are possible for us to do (see Table 3.4). We will get further into this in Section 3.6 describing the experiment design of the DMOS test.

3.3 DMOS Test Recruitment and Practical Considerations

According to the P.800 recommendation the environmental noise for the listening test should be kept as low as possible, setting a recommended limit below 30 dBA [5]. It also states that the room size should be between 30 and 120 cubic meters and have a reverberation time of less than 500 ms. We have decided to deviate slightly from these recommendations for a few reasons; we do not have immediate access to the equipment to measure these things, nor do we have access a professional recording studio that may ensure some of these requirements. It would also add a fair bit of complexity to our already quite broad study in terms of different activities needed to be done to acquire these things. This would take time and re-

sources in itself, but it would also make it more practically difficult to conduct our subjective DMOS test. One of the reasons for this is that the participants would have to travel to our recording studio, and/or having to rent/borrow a studio and the necessary measurement equipment at the same time as the participants also have time to participate would add complexity to the time management of our study. This could likely result in the recruitment of less participants, as it would be more work for them to travel to us and they may also not be free when we have access to the studio and equipment needed.

Because of this we decide to go with a more practical methodology that we still think covers the ideas behind the P.800's recommendations [5] to a high degree. Instead of using these very rigid room requirements, we travel to our participants preferred locations to conduct our subjective DMOS tests. At these locations we look for the quietest room available with the least distractions from other people. All participants are also wearing a pair of Sony WH-1000XM3 Noise canceling headphones with noise canceling enabled during the DMOS test. As mentioned, we still look for the quietest rooms possible as the noise canceling headphones are not perfect and may let in some noise or otherwise alter the audio in noisy environments. We believe this ensures enough consistency between our different test environments to ensure valid and comparable results.

Since we are traveling to the participants the audio samples are played from a laptop. The same Lenovo Thinkpad L13 is used for all of our testing to ensure consistency. We also make use of a 3.5mm audio cable for the noise canceling headphones, in order to remove any potential audio disturbances or quality issues that may arise from using Bluetooth. More technical information about the experiment can be found in Section 3.6.2 explaining the experiment setup and technical details.

To ensure proper hygiene and that the participants are comfortable with participating, single-use disposable ear-cup covers are used on the noise canceling headphones. These are of course replaced for each participant. To ensure that the participants are comfortable they are also allowed to pick a comfortable volume themselves before the DMOS test. They are however, instructed to pick the loudest comfortable volume. The reasoning behind this is further explained in Section 3.6 on the experiment design, along with more details.

Participants are mainly recruited from family, friends and their acquaintances. If more participants would still be needed after exhausting these alternatives there was a backup plan in place to recruit students and employees at NTNU, in addition to recruitment from social media channels. However, this has not been necessary as 21 participants are recruited from the first group. The initial recruitment goal was 20 participants, as this is a number that we have seen used in several other studies [4, 16], and one that seemed realistic considering the time frame of the thesis. [16] achieved close correlation between MOS, DMOS and PESQ with this number of participants. [4] appear to have used the older, narrow-band version of PESQ as they state that the PESQ scores are given between 1 and 4.5, they achieve a subjective MOS score of 5 and a PESQ score of 4.47 while testing their

proposed audio steganography method. It was difficult to find an exact recommended number of participants for audio or speech quality MOS testing, but this paper on MOS limitations [55] refers to the audiovisual MOS testing ITU-T recommendation P.911's [56] recommended participant number of anywhere from to 6 and 40 and the video MOS testing ITU-T recommendation BT.500's [57] stated recommended minimum of 15 participants. This survey investigating how differences in methodology affects MOS scores for TTS evaluation uses 26 participants for each of its four different MOS methodologies [58]. Judging from these papers 20 participants seems like a fair number for our study.

One extra participant is included as they were practical to include at the time of conducting the DMOS test, and because we deem it advantageous to have more participants than required in case of any outliers being excluded from the study. Assuming some potential differences in DMOS ratings given by people of different ages and genders, some efforts are also made to try to recruit participants from different gender and age groups. However, we assess it to be more important to get enough participants than maintaining this balance, so we do not let this consideration stop us from recruiting willing participants. The age groups and genders of all participants are collected and the participant demographics are reported in Section 4.1.1.

During the pre-project [9] it was identified that any study at NTNU that collects personal information needs to be registered with Sikt, which provides privacy services for the university [59]. It was not clear to us if the age and gender information we planned to collect classified as personal information, so Sikt was contacted to check this. They encouraged us to register the study with them for approval [60], which was done shortly after handing in the pre-project report in December 2024. This also meant we had to create a consent form to be signed by all participants of our Subjective DMOS test. This was done using Sikt's template [61] and can be seen in Appendix C. The project was swiftly approved within a few days after reporting it to Sikt.

3.4 Cover File Selection

3.4.1 Dataset Selection

Our dataset is chosen in adherence to the P.800 recommendations [5]. Both the ACR and DCR procedures of the P.800 are deemed relevant for our dataset selection. This is because the DCR procedure explained in the P.800 is not nearly as detailed as the ACR procedure, especially when it comes to details concerning the selection of audio samples. For instance, the DCR method does not state the desired length of the chosen sentences, nor does it specify the use of speakers with different genders. It does however mention that the DCR method is largely based on the ACR procedure, as well as explicitly mentioning parts of the ACR method that does not need to be followed in DCR, such as accounting for the order of presentation effect. We therefore decide to follow the P.800 ACR method's recom-

mendations relevant to dataset selection for the areas where the DCR procedure did not explicitly state otherwise or have its own recommendations.

The P.800 ACR recommendation [5] stresses the importance of including both male and female speakers in the tested speech samples. The ACR procedure recommendation also emphasizes the importance of using more than one speaker of each sex to avoid having the results potentially skewed by peculiarities of a single individual's voice. The ACR procedure recommendations also state that the recordings used for the MOS test should be of studio quality. The DCR specific recommendations recommends the inclusion of at least four speakers, further stating that the speakers should all be reading the same two phonetically rich sentences during the test. In addition to following the relevant parts of the P.800 recommendation, we add the requirement of the dataset being chosen for this study being free of charge for academic use.

As discovered during the research project planning course [9] planning out this study, the TIMIT dataset [36] stands out as a possible contender, as Reyers [3] evaluated this to be suitable for the audio steganography tools and methods that we will be using in this project, in his thesis [3] comparing these methods. Reyers did this by comparing four datasets identified as the most commonly used ones within the field in a review of the current state of the art in audio steganography from 2020 by AlShabany et. al. [2] and his reasoning is compelling. However, on further inspection it is discovered that the TIMIT dataset does not currently appear to be free to use for academic purposes [36].

Alternative datasets are therefore found with assistance from the Microsoft Copilot [20] AI assistant. Copilot is prompted to suggest alternative datasets similar to TIMIT [36], that also follows the requirements from the P.800 mentioned earlier. The suggestions are then reviewed manually. After looking through many different proposed datasets we find three that appear to be possible contenders: The "MOCHA MultiCHannel Articulatory database" [62] (MOCHA-TIMIT), the "Pitch Tracking Database from Graz University of Technology" [63] (PTDB-TUG) and the "CMU ARCTIC databases for speech synthesis" [64]. All of these datasets appear to be recorded in professional recording studios with the PTDB-TUG dataset containing phonetically rich sentences, and the two others containing phonetically balanced sentences.

The PTDB-TUG dataset initially seems like a strong contender as it contains ten speakers of each sex and only contains phonetically rich sentences [63]. In addition to this the PTDB-TUG dataset is also newer as it was released in 2011 compared to the MOCHA-TIMIT and CMU Arctic datasets that were released in 1999 and 2003 respectively [62, 64]. Which we initially thought could mean it had higher quality recordings. However, on further inspection it is noticed that the ten speakers do not all speak the same sentences and that it is not possible to choose four identical sentences spoken by each of our needed two male and two female speakers. Many of the sentences also contained breathing noises and button presses, and while some of them only contained these noises before and after the sentence was spoken so that they could be edited out somewhat easily, this was

not the case for all of the sentences. This would complicate both the audio selection process as it isn't always easy to hear whether the breathing or button noises can be easily edited out or not. It would also require an extra pre-processing step for editing out these noises before the audio files would be ready for embedding which would take time. The MOCHA-TIMIT dataset is also excluded for similar reasons as it only contains one speaker of each sex, which is not enough for our requirements.

Ultimately, it is discovered that the CMU ARCTIC dataset contains a more than sufficient amount of male and female speakers reading out the same sentences [64]. The four speakers labeled as US English bdl (male), US English slt (female), US English clb (female) and US English rms (male) are chosen from the main CMU ARCTIC database as they speak clearly and have recordings of pretty good quality. The sentences also appear to have been edited to not contain much space before and after the spoken words, as well as seemingly having the least amount of breathing or other unwanted noises in its recordings. This comes at the cost of the microphone quality not always being the best, with some unwanted noises seemingly coming from the microphones being present in some of the recordings. The dataset also contains phonetically balanced sentences, rather than phonetically rich ones. However, this was the only dataset found that otherwise fit all of our required criteria. At this stage a considerable amount of time has also been spent on finding a suitable dataset, and this one will therefore have to be selected in order to ensure ample progress of the project.

3.4.2 Audio Selection

The P.800 DCR and ACR procedure recommendations [5] are also used when selecting our audio samples for our study from the CMU ARCTIC dataset [64]. The P.800 recommendations do not mention any audio file length for the DCR method specifically. The recommendations for the ACR method of about two to three seconds per recorded sentence is therefore followed [5]. It makes sense to use the ACR method recommendations for the sentence length as both PESQ and ViSQOL appear to have been developed with this type of Subjective MOS test in mind. This is explicitly stated for PESQ in [43] and suggested by comparison to a Subjective ACR MOS test in Experiment 1 of the paper proposing ViSQOL as well as being mentioned as being the most common subjective MOS methodology in the background section of the same paper [12]. It therefore seems rational to us to follow this recommendation in order to increase the chances of the audio samples working well with our MOS-LQO algorithms, as the same samples will have to be used for both our subjective and objective MOS testing in order to compare the algorithms.

The P.800 ACR specific recommendations also mention that a windscreen should be used if breath puffs from the speakers can be noticed [5]. While it is possible that this may not affect our DMOS results, as it is not explicitly mentioned in the DCR specific recommendations of the P.800, we imagine that it could at the very

least affect the results of our MOS-LQO testing. We therefore take this into consideration while choosing our audio files by trying to select audio files with minimal "non-speech" breathing or mouth sounds. From this we also assume that other "non-speech" disturbances like button clicks or background noise could possibly also affect our MOS-LQO and DMOS test results negatively. We therefore try to avoid audios that contain any non-speech sounds to the best of our ability when selecting our audio files.

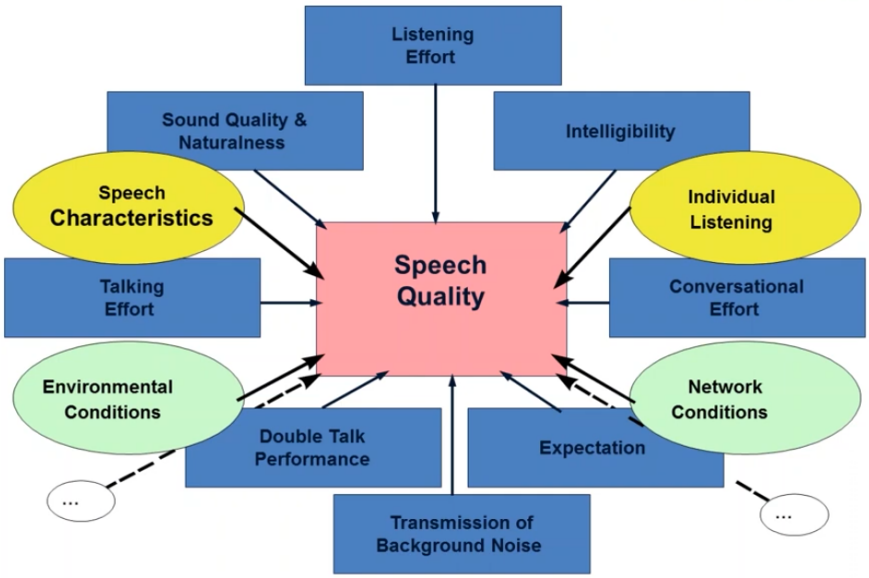


Figure 3.1: Different things that can affect speech quality, adapted from [65].

Effects mentioned in a seminar of mean opinion score testing of voice quality hosted on HEAD acoustics International's, an acoustics company that has been in the industry for 40+ years, YouTube channel and presented by Jabob Sondergaard is also taken into consideration during the audio selection process [65]. The seminar showcases different effects than can affect speech quality MOS scores. Even though we are interested in audio quality rather than speech quality, we think that it is possible that many of these effects could also affect how participants rate audio quality. We therefore deem the effects mentioned in this seminar to be relevant to our audio selection, as we do not want external factors to affect our MOS scores, but rather want the effects measured to be isolated to degradations in audio introduced by our audio steganography methods. We think this will allow us to more effectively evaluate what MOS-LQO algorithm is most suited to measure the perceptual transparency of audio steganography methods. The different effects showcased in the seminar can be seen in Figure 3.1.

Two effects highlighted by the HEAD seminar [65] early on strikes us as particularly relevant for our study. While many of the other effects appear to either not be applicable to us or otherwise already being taken care of, for example by

following the P.800 standard or using audio files recorded in a recording studio. Sondergaard refers to these two particularly relevant effects as intelligibility and naturalness. While intelligibility has a quite obvious definition as how comprehensible speech is [66], Sondergaard states that naturalness is bit of a vaguer concept [65]. Sondergaard explains speech with low naturalness as speech that might be perfectly intelligible, but that we still wouldn't say sounds natural. He uses Stephen Hawking's synthesized speech as an example of this [65].

These concepts are applied during our audio selection process by excluding sentences that for any reason sounds unnatural to us from being selected. This can for instance be sentences that have unconventional sounding connotations of specific words, or sentences that have unnatural pauses, i.e. sentences that do not follow a natural "flow". The intelligibility effect is taken care of by excluding any sentences that we find can be difficult to understand from the selection. If this is true for any of our chosen speakers in the dataset, another sentence needs to be selected for all speakers as they all need to repeat the same four sentences in accordance to the P.800 DMOS procedure [5].

The audio files are picked by listening carefully to several recorded sentences from different speakers from the CMU ARCTIC dataset [64], using the same pair of headphones that will be used by the participants of our Subjective DMOS test. The two female and two male speakers that we evaluate as generally sounding the clearest to us are then selected. Four sentences that are close to two to three seconds in length, and that sound clear and good across all of the selected speakers, as well as fulfilling our requirements for intelligibility and naturalness, are then chosen. We also try to exclude sentences containing any potentially sensitive topics that could possibly disturb our participants. For instance, one sentence containing references to weapons is excluded on these grounds.

The sentences labeled "a0011", "a0069", "a0094", "a0154" are selected from the CMU ARCTIC dataset [64] by following the procedure described above. The spoken contents of these sentences can be seen in Table 3.2. These four sentences are later combined in pairs divided by 0.5 seconds of silence to make up two so called "audio samples" for each speaker, in accordance to the P.800 DCR recommendations [5], following the order that they are presented in above. More information about the presentation of the audio files and an illustration showing the silence dividing the audio files and audio samples can be seen in Section 3.6 showcasing the Experiment Design of the Subjective DMOS test. The chosen audio files are all 16-bit single channel (mono) PCM WAV files with a sample rate of 16.000 kHz and bit rate of 256 kbps.

The selected audio files can be found on our GitHub [67] in the "Selected Audio Files" directory in "Supplementary Materials".

3.5 Applying Audio Steganography Methods

After selecting our audio files we can start implementing and applying our chosen audio steganography methods to them, before eventually combining the files into

File name	Sentence spoken
a0011	If I ever needed a fighter in my life I need one now.
a0069	It was his intention to return to Eileen and her father.
a0094	He had barely entered this when he saw the glow of a fire.
a0154	He was smooth shaven and his hair and eyes were black.

Table 3.2: The sentences spoken in the the selected audio files from the CMU Arctic [64] dataset.

audio samples for our MOS tests. This section explains the implementation and application of each audio steganography method. As mentioned in Section 2.3 our steganography tools and methods and chosen to be the same as were used in Reyers thesis [3], Section 2.3 also explains our reasoning for this.

In Reyer's thesis, it is reported that for Steghide and Hide4PGP higher embedding capacities lead to worse SNR scorer. Therefore, we have chosen a methodology to embed close to maximum capacity into these methods in order to create a worst case scenario. For the GAN method, Reyers reports an almost flat SNR score across different embedding capacities. We find this a bit strange and therefore apply this method at two hiding capacities, one near maximum and one of about 5% to see if we get similar results. As mentioned previously in this thesis, we refer to the GAN method applied at these different capacities as GAN low and GAN high. The TAN method is implemented, but not applied to all of our selected audio files. This is because we had trouble implementing it in time for the subjective DMOS test. However, we were eventually able to implement it after finding a mathematical error in the paper [19] proposing it. More about this can be seen in Sections 3.5.4 and 4.5.

The subsections in this section show the details of how our different chosen audio steganography tools and methods are implemented and applied.

3.5.1 Steghide Methodology

For the Steghide methodology a Kali Linux 2024.4 VirtualBox virtual machine (VM) image installed from the official Kali website [68] is used with Oracle VirtualBox [69] Version 7.1.6 r167084 (Qt6.5.3). Steghide [30] version 0.5.1 is then downloaded from the default official Kali Linux repository with the apt command.

To create the Steghide stego files the maximum capacity of each cover file is first checked with the built in Steghide "info" command. This command returns the maximum capacity of each file in kilobytes (KB), which is then converted to bytes using this online converter [70] (the binary result is used), before using this online Lorem Ipsum generator [71] to generate a random Lorem Ipsum text that is exactly the size of the maximum embedding capacity of each file to be used as the secret message. This generated text is then copied for each cover file and saved as txt files to be used for embedding (Figure 3.2 illustrates the process up to this point), before embedding each txt file into their respective cover files with

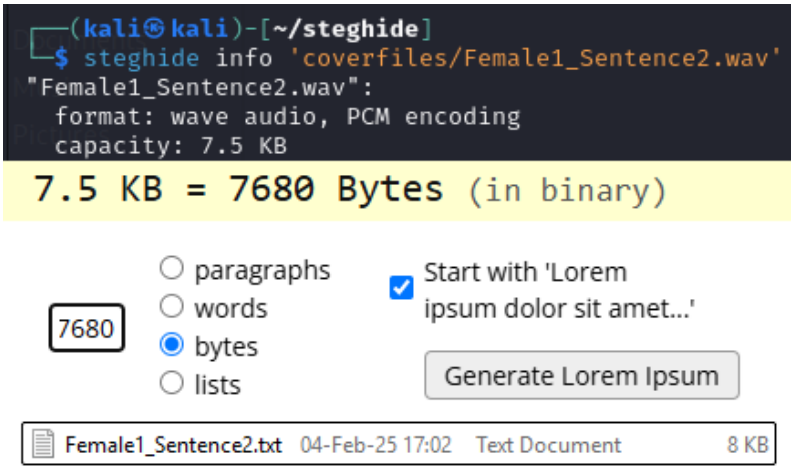


Figure 3.2: How the Steghide maximum capacity secret message is created.

the following command:

```
steghide embed -ef 'secrets/[SECRETMESSAGE].txt'
-cf 'stegofiles/[STEGOFILE].wav' -p password123
```

The command assumes that all the cover files have been copied over to a directory called "stegofiles" before it is ran. This is done so that the original cover files won't be altered by the command since it changes the specified files directly instead of copying them to a new directory and keeping the originals. "embed" specifies to Steghide that we want to embed information into a file, "-ef" specifies the path of the file to be embedded into the cover file, "-cf" specifies the path of the cover file, and -p specifies a password needed to recover the message from the stego file again after embedding, which is set to "password123" for all of our cover files.

To make sure everything went as expected we also extract the information from each stego file and compare the extracted message txt file to the original embedded txt file manually to ensure that they are the same. The extraction is done using the command:

```
steghide extract -sf 'stegofiles/[STEGOFILE].wav'
-p password123 -xf 'extracted/[EXTRACTEDSECRET].txt"
```

This process is repeated for all 16 cover files selected in Section 3.4 (four files each for four speakers) to create stego files that are later used to make audio samples for our subjective and objective DMOS, MOS-LQO and SNR testing. The stego files retain all the same file properties in terms of sample rate, bit rate, etc. as explained in Section 3.4.2.

We decided against automating this process, both because the time saved would be limited, and because we didn't want to risk bugs in the automation

code potentially messing up the embedding process. We also considered the possibility of these bugs staying undiscovered because of worsened oversight of the embedding procedure.

3.5.2 Hide4PGP Methodology

The Hide4PGP methodology shares many similarities with the Steghide methodology, but is slightly more complicated because of its PGP integration. Hide4PGP 2.0 downloaded from [32] for Linux is used on an Ubuntu 22.04.2 LTS virtual machine (VM) running on Microsoft's Windows Subsystem for Linux (WSL) 2 [72]. WSL 2 was chosen in favor of VirtualBox [69] this time around as it appears to run faster and overall give us less problems. However, we did get some permission issues with Ubuntu and decided to fix these quickly by simply doing all commands as the root user, as we were only going to use this VM for our Hide4PGP stego file creation.

Like with Steghide, we start by checking the hiding capacity of each file, which is done with the "-i" parameter in Hide4PGP. An example of how this command can look is shown below:

```
./hide4pgp -i "coverfiles/[COVERFILE].wav"
```

"./" is used in front of "hide4pgp" as it is ran straight from its 32-bit executable. "-i" specifies that Hide4PGP should show information about the cover file, such as the file type and hiding capacity. However, we cannot just generate Lorem Ipsum text using the hiding capacity displayed here like we did with Steghide. This is because the Hide4PGP documentation included with the install [32] states that the secret message should be encrypted with PGP before embedding, and in our experience this process slightly increases the file size of the secret message. Because of this, we add a "safety margin" of one kilobyte (1024 bytes) to make sure that all the PGP-encrypted secret messages will embed successfully. This is done by subtracting 1024 bytes from the maximum embedding capacity shown by Hide4PGP, before generating the Lorem Ipsum with the same Lorem Ipsum text generator used in the Steghide methodology [71]. Like Steghide, Hide4PGP also reports maximum hiding capacities in kilobytes, so the same online kilobyte to byte converter that was used for Steghide is also used for Hide4PGP, before subtracting one kilobyte and generating the secret Lorem Impsum message.

For each file, this generated message is put into a txt file which is encrypted with PGP using RSA-2048, by utilizing the GnuPG (gpg2) 2.4.4 program typically included with Ubuntu. For ease of use, we initially use the GPG file format rather than the older PGP format. However, we later noticed that we had been extracting the files in the PGP format, but this made no difference in being able to recover and decrypt the secret message. This is also reflected in the showcased commands.

The generated secret message is first encrypted with gpg2 by using the following command:

```
gpg2 --encrypt --recipient root --output
./encrypted/[ENCRYPTEDSECRET].gpg
./secrets/[PLAINTEXTSECRET].txt
```

The encrypted secret message is then embedded into each cover file to create our Hide4PGP stego files with the following command:

```
./hide4pgp stegofiles/[STEGOFILE].wav
encrypted/[ENCRYPTEDSECRET].gpg
```

The command above assumes that all cover files have been copied over to a directory called "stegofiles" before it's ran. This is done so that the original cover files won't be altered by the command.

Each secret message is also extracted from the stego files, decrypted and manually compared to their original embedded counterparts to ensure that the message is recoverable and unaltered. The extraction is done using the following command:

```
./hide4pgp -x stegofiles/[STEGOFILE].wav
extracted/[EXTRACEDSECRET].pgp
```

The "-x" specifies that Hide4PGP should extract a message from the stego file.

Lastly, the extracted message is decrypted with gpg before it can be compared with the original secret message, using the following command:

```
gpg2 --decrypt --output ./decrypted/[DECRYPTEDSECRET].txt
./extracted/[EXTRACTEDSECRET].pgp
```

This process is repeated for our 16 selected cover files to create our Hide4PGP stego files that are later used to create the audio samples for our subjective and objective MOS testing. The stego files retain the same file properties in terms of sample rate, bit rate, etc. as explained in Section 3.4.2.

3.5.3 GAN Based Audio Steganography Implementation and Method

The GAN method implementation [39] used in this thesis is an extension and implementation of the two papers [37] and [38]. This implementation supports both image and audio steganography. The implementation is largely left unchanged, however extensive work is put into to making it work by installing the correct versions of dependencies, as is typical for getting older Python projects to work. Python 3.6 is chosen to be the most likely contender to have been used during the implementations development judging from the initial creation date of the GitHub project [39], and this version is therefore used to run the implementation. Versions needed to be specified for most dependencies in order to get the code to run, a combination of trial and error and asking Microsoft CoPilot AI [20] about what dependency versions work together, and that may fix various error

messages, is used to eventually specify and install a working combination of dependency versions.

All our cover files were manually resampled from 16.000 kHz 44.100 kHz with Audacity [73] before embedding. The report from the GAN implementation authors [39] also posted on the GitHub states to use 5 kHz or 22 kHz sample rates, however when resampling our cover files to 22.100 kHz, the stego file generated by the GAN method appear to play back in double speed. Inspecting the stego file's properties reveals a sample rate of 44.100 kHz which is why we decided to try resampling our cover files to this sample rate instead. Doing this resolves the double speed problem and the samples appear to play back at a normal speed.

The same pre-trained model that was used by the authors of [39] in their Python Jupyter Notebook for applying the audio steganography part of the GAN method was used also for our purposes. This may not be optimal as the authors state that this was trained mostly on music, while we are operating on audio files containing spoken english. This does appear to make a slight difference as Reyers [3] report a slightly better SNR score for their tested music samples than their spoken english samples when applying the GAN method in their thesis. However, the SNR scores for both their spoken english and music datasets are both very poor and seem to get closer to each other at maximum capacity. We also measure way higher SNR scores than Reyers results in our experiments, is is possible that they used one of the recommended sample rates or another one of the included models. We also don't know for sure if Reyers used the same GAN implementation as us, as they only refer to the original GAN paper [37] in their thesis. Since we are mainly interested in evaluating ways to measure perceptual transparency, rather than the audio steganography methods themselves, we will not be spending time on training our own model for the GAN method. In fact, having a method with worse performance could even be beneficial in order to test a wider range of degradations.

As mentioned in Section 2.3.3 about the GAN method's background, we divide the GAN method into two hiding capacities which we refer to as GAN High and GAN Low. GAN High is the GAN method applied at its maximum capacity for each file, while GAN Low is the GAN method applied at 5% of maximum capacity for each file. Unlike Steghide and Hide4PGP, the GAN method does not have a way of showing the maximum hiding capacity of a file. We therefore decide that this will have to be brute-forced by embedding larger and larger messages. Several attempts are made to automate this process. One attempt adds one by one letter to the input until the embedding fails, saving the last input and one attempt does the same with extra words instead. However, both of these methods encounter strange errors where the embedding fails on inputs way smaller than the maximum capacity. It appears to us that certain words or characters can sometimes get the embedding to fail, even for inputs much smaller than the max capacity. A manual method is therefore ultimately chosen, as we do not have time to figure out exactly what is causing these strange errors.

The manual method chosen is similar to the Steghide and Hide4PGP meth-

ods, but differs in some areas. For the GAN high method a plenty large enough Lorem Ipsum text is generated with an online Lorem Ipsum generator [71]. Once this fails to embed, parts of it are manually removed until it no longer fails. Then, letter by letter is added back to find the exact maximum capacity of the file. When this is found the GAN High stego file can be generated with a maximum capacity message embedded. After this is done, characters are counted to determine the amount of bytes embedded, as all normal English capital and non-capital characters, as well as commas and punctuation marks, take up one byte in Python UTF-8 encoded strings. After counting the amount of characters/bytes in the maximum capacity message we can multiply the resulting size by 0.05 to find 5% of maximum capacity for the current cover file. After doing this, characters are removed to leave just the resulting 5% of max byte size and the GAN Low stego file is generated. This is repeated for all cover files to generate our 16 GAN High and 16 GAN Low stego files. The stego files generated by the GAN method are 44.100 kHz 32-bit float wav files with a bit rate of 1411 kbps.

The final code used for the GAN method implementation can be found on our GitHub in the "GAN Method Code" directory [67].

3.5.4 TAN Based Audio Steganography Implementation

The TAN method proposed in the paper [19] was implemented from scratch with the help of Microsoft CoPilot [20] and OpenAI ChatGPT [74]. This method contains three main components: AES-128 encryption, a two dimensional logistic tan map (2D TAN map), and code for embedding and extracting the secret message using the outputs of the TAN map. The AES-128 implementation was largely generated by CoPilot, while the rest of the code was largely generated and modified by both ChatGPT and CoPilot. Naturally, some manual work was also done to make everything work together, as well as fixing some bugs in the AI generated code.

The TAN method uses a least significant bit (LSB) approach where an audio file is split up into its (in our case 16-bit) audio samples, and the deterministic (for the same starting input) chaotic output generated by the 2D TAN map is used to decide what order of samples LSB's are used for embedding the secret message.

II. THE 2D LOGISTIC-SINE MAP (2D-LSM)

The 2D Logistic-Sine map (2D-LSM) is defined by Eqn. (1)

$$\begin{cases} x_{n+1} = \sin(\pi a(y_n + 3)x_n(1 - x_n)) \\ y_{n+1} = \sin(\pi a(x_{n+1} + 3)y_n(1 - y_n)) \end{cases} \quad (1)$$

where a is a control parameter and $a \in [0, 1]$, x_n, y_n are output values.

B. 2D Logistic Tan Map

The 2D Logistic Tan Map is defined by

$$x_{n+1} = tan((\pi r y_n + 3)(x_n(1 - x_n))) \quad (1)$$

and

$$y_{n+1} = tan((\pi r x_n + 3)(y_n(1 - y_n))) \quad (2)$$

where r is a control parameter and $r \in [0, 1]$, x_n, y_n are the output sequence values used in the key. The 2D Tan Logistic

Figure 3.3: The sine map functions from paper [22] and the tan map functions from paper [75]. Adapted from [22] and [19].

The AES-128 encryption component was developed by the help of Microsoft CoPilot [20] and was quite straight forward to implement. The 2D TAN map defined in the paper [19] however, is not so trivial to implement, even though Microsoft CoPilot was also used to develop this component of the stego method.

The 2D TAN map is first implemented exactly as it is described in the paper. To confirm that we have the correct implementation we try to replicate the output values of the 2D TAN map implementation used in the TAN paper [19] and showcased in Table 1 of their paper. An adapted version of their table can be found to the left of our Table 3.3. To do this, we need to replicate the tan paper's starting values for x and y, and the r parameter as required by the tan map functions shown to the right of Figure 3.3. The starting values for x and y are both shown to be 0.3 in the tan paper, but they do not specify the r value used. To find the r value values from 0.1 to 0.9 are tried in increments of 0.1 and the output is inspected. When using an r value of 0.9 it is observed that the same output as the first iteration of the tan paper's [19] TAN map output can be reproduced, but after the first iteration the values all deviate. Since the first values being both being identical by chance seems unlikely, this prompts us to look for any potential mistakes in the TAN paper.

Output by TAN paper (x, y)	Output by our implementation (x, y)
(10465258, 30235701)	(10465258, 30235701)
(7346647, 6264212)	(7346647, 6264212)
(13410269, 26680129)	(13410269, 26680129)
(5695430, 102796485)	(5695430, 102796485)
(52553839, 14999801)	(52553839, 1431521)
N. A.	(14999801, 12300396)

Table 3.3: The 2D TAN map output from the TAN method paper [19] compared to the output of our implementation, partly adapted from [19].

Eventually, an error is found in the TAN paper [19] by looking at the papers cited and discovering a sine map steganography method paper [22] by the same authors [19] that cites yet another similar sine map paper [21] which both propose very similar methods to the TAN paper, using a sine map instead of a tan map. The y function in the last of the two mentioned sine papers [21] contains x_{n+1} in its y function to offset the iterations of x and y between the two functions, while the TAN method paper's y function simply reads x_n, meaning that the values are not offset. This difference is showcased in Figure 3.3 with the "+1" omitted by the tan paper marked in red.

When adding this iteration offset to our implementation of the TAN method, we recognized many of the same output values. However, we noticed that we also had a many output negative values that were not shown in the TAN paper. We therefore concluded that the TAN papers implementation likely simply got rid of all negative values and used every other positive value in the order they were left behind as either x or y values, regardless of what function created them. After

implementing this logic we are able to produce a nearly identical output to the TAN paper.

Our TAN map's outputs compared with the tan paper's outputs can be seen in Table 3.3. The only difference between the two can be seen in iteration five, where the TAN paper reported a y value of 14999801, and our implementation produced a y value of 143121. However, the next output value produced is 14999801 for our implementation. We would argue that this strongly suggests that this is another spelling mistake by the TAN paper, as we find it very unlikely that they produced all the same output values as us, before somehow skipping one value and going back to produce the same next output value.

Finally, the code for embedding and extracting the message by splitting up the audio files into their (in our case 16-bit) audio samples, and changing their LSB's in the order generated by the TAN map, and in the way defined by the TAN paper, was created with help from Microsoft CoPilot [20]. All the code, including a proof of work demonstration of the entire method as proposed in the tan paper [19], applying AES-128 to a secret message before embedding it in an audio file, and doing the whole process in reverse to extract it again can be found on our GitHub in the "TAN Method Code" directory [67].

As mentioned, we did not have time to create stego files using the TAN method before our first appointment with test subjects for the objective DMOS test, because of the errors discovered in the TAN paper. We therefore decided to include two capacities of GAN instead of the TAN method in the DMOS and MOS-LQO tests, as described in Section 3.5.3, and report the discovered TAN errors and solutions as an additional result.

3.6 DMOS Experiment Design

After choosing the DCR method for our subjective MOS test in Section 3.2, we can go ahead with the experiment design of our Subjective DMOS test based on the P.800 recommendation [5]. The P.800 annex for ACR refers to annex A for conversation opinion tests which names a number of experiment types that can be used, such as Latin Squares [76] and Randomization with Replication [77], but ultimately states that the researcher needs to decide what is best for their own use case [5]. The P.800 ACR procedure also specifies that the order of presentation effect needs to be accounted for, which is something that can intuitively be done with some sort of simple randomization of order of steganography method for each participant or for instance by the use of Balanced Latin squares [78]. However, doing this would add complexity for our manual approach using pen and paper, and while it would be possible to implement in our experiment design it would also leave more room for manual errors during it's conduction. Luckily, the P.800 DCR procedure we have chosen also explicitly states that the order of presentation effects does not need to be accounted for [5]. We therefore decide against using any of these experiment types mentioned by the ACR procedure and instead design our own experiment that we think will allow us to compare how

good our chosen MOS-LQO Algorithms are at evaluating the perceptual transparency of our chosen audio steganography methods compared to Subjective DMOS in a good way.

We are taking a manual approach to our experiment and this is done to ensure that we have time to conduct it with the somewhat limited resources of a single person's Master's thesis. We do not find it appropriate to spend time on developing or investigating potentially already existing frameworks for our test, as we are already under time pressure from all the activities needed to be done for this thesis. We therefore decide to make use of a experiment design where the audio is manually controlled by the person conducting the study and the ratings from the participants are collected with pen and paper. Which we will now explain the details of in greater detail in Section 3.6.1.

3.6.1 Experiment Procedure and Information Given to Participants

First and foremost, the participants are given some information about the study and the law obliged consent form displayed in Appendix C is read and signed by each participant. More detailed information about the practicalities of the study are given to the participants in the form of printed paper displayed in Appendix D, as is recommended by the P.800 [5]. The first three participants are given the same information as contained in the information paper verbally before discovering that an information paper might be more effective for this purpose. Some of the participants were also sent this information by E-Mail before the test took place in order to save time by allowing them to read it beforehand. The participants are then given a chance to ask practical questions about the test and are told that they can also ask questions during the test if something is unclear. The P.800 ACR recommendations are followed when it comes the types of questions that are answered, answering questions about the meaning of the instructions given and general questions about the procedure, but not technical questions [5]. For instance, the participants are not informed about the existence of a null pair even if they ask about this directly.

In accordance to the P.800 [5] ACR procedures participant requirements, each participant is asked whether they have; "...been directly involved in work connected with the assessment of the performance of telephone circuits, or related work such as speech coding" [5, p. 18], "not participated in any subjective test whatever for at least the previous six months, and not in any listening-opinion test for at least one year" [5, p. 18], and have never heard the sentences presented in the tests audio samples before. All the participants answer no to all of these questions. The P.800 ACR participant requirements are used here as there are no explicit participant requirements mentioned in the DCR procedure.

The consent form in Appendix C given to the participants contains some general information about the experiment and its purpose, along with some privacy related information about how their data is processed etc. The information paper in Appendix D given to the participants contains more detailed information about

5	Degradation is inaudible. *Nedgang i lydkvalitet er ikke hørbar.*
4	Degradation is audible but not annoying. *Nedgang i lydkvalitet er hørbar, men ikke irriterende.*
3	Degradation is slightly annoying. *Nedgang i lydkvalitet er litt irriterende.*
2	Degradation is annoying. *Nedgang i lydkvalitet er irriterende.*
1	Degradation is very annoying. *Nedgang i lydkvalitet er veldig irriterende.*

5 ---------- 4 ---------- 3 ---------- 2 ---------- 1

<-Least annoying (better quality) ---------- Most annoying (worse quality)->

<-Minst irriterende (bedre kvalitet) ***Mest irriterende (værre kvalitet) ->***

Figure 3.4: The DMOS Scale explanation and translation used in the participant information paper.

the actual procedure of the test. The document explains that the participants are to rate the *degradation in audio quality* between some different degraded audio samples and their high quality reference on a scale of 1-5. The scale explanation used in the document can be seen in Figure 3.4. This scale also includes a translation of the P.800 DCR scale [5] to Norwegian as the study is conducted in Norway. This translation is made by us and great care is put into making it as direct of a translation as possible. The information paper also includes a simplified illustration of the papers that the participants will be given to mark their ratings on, this simplified version can be viewed in the original information paper in Appendix D.

Figure 3.5 displays the first page of the two page paper given for each degradation to collect the ratings in the actual test. The first page contains the ratings to be given for the four female audio samples for each method, while page two contains the ratings to be given for the four male audio samples. Two pages like this are given for each of our four stego methods as well as for our null-pairs, so each participants fills our ten pages in total. The rating paper shows each tested steganography method and null-pair method as one "Round" so that the rating for each method can be correctly interpreted later.

The null-pair is included in accordance to the P.800 DCR procedure recommendations in order to be able to check the quality of the anchoring [5]. In the null-pairs the high quality reference is also played as the "degradation", instead of a sample altered by a stego method. The round numbers on the rating papers are already filled in by us along with the participant number before handing the papers to the participants. The participants age, sex and chosen volume is filled in by us after they have completed the test. The participant number is used as an anonymous identifier for each participant and the chosen volume is collected to have better grounds for excluding outliers if they have selected a different volume level than most of the participants, and have given ratings deviating from the norm.

Volume is quickly discovered to seemingly make a big difference for some participants. When we refer to volume levels in this thesis we are referring to Windows 11's built in volume adjustment. Initially, the participants are allowed to pick any volume they want by the use of the scroll wheel on a wireless mouse. The volume is started at 50% and the participants are instructed to scroll up or down from here. This is done by playing one sentence from each of our selected speakers from the CMU Arctic dataset [64] that are not used as samples in our actual DMOS test. This method is chosen to ensure that the participants have a comfortable volume across all of the different speakers. The logic behind letting the participants pick their own volume is that hearing is subjective, and we expect that the same volume level can be perceived differently by our different participants. This logic is also backed up by the P.800 ACR procedure where it's stated that "there is no universal optimum listening level" [5, p. 16]. Therefore, we do not think that forcing every participant to use the same volume will add consistency to the test as this volume might be too loud or too quiet for some participants to effectively percept the degradations of audio quality.

However, this first method for picking the volume quickly appears to not have

been defined strictly enough as one of our first three participants chose a Windows 11 volume level of 12% while the two others both chose 100%. The participant selecting 12 as the volume gave both the GAN High and GAN Low methods the same DMOS scores of 4.875, while the other participants gave them 3.5, 3.7, 1 and 1. This participant was therefore asked to try to listen to the samples again with higher volume, which they did. They then expressed that they would have given the samples a way lower rating if they had used a higher volume, and that they could have probably quite comfortably have used a higher volume for the test. The participant giving these high ratings were therefore excluded completely from the test as an outlier. Z-scores were later also used to statistically confirm them as an outlier, which also resulted in the participant giving the ratings of 3.5 and 3.7 being excluded. These statistics are explained in further detail in Section 3.9.1 on excluding outliers.

Because of this early volume method observation, the way of deciding the volume for each participant is changed. With the new method being more strictly defined and rigid to try to prevent participants from picking a volume that is so low that it affects the results significantly. In the new method, the participants are instructed to pick the loudest volume that they can possibly be comfortable with. The same volume test samples are used, but the participants are now made to listen to the max volume first and are instructed to verbally tell the conductor of the experiment if the volume is too loud. If this is the case the volume will be adjusted down in intervals of 25%. This gives a total of four possible volume levels: 100%, 75%, 50% and 25%. The reasoning behind using only four steps is that it divides participants into fewer volume groups where statistics can be used more effectively to check if the difference in volume lead to a significant change in DMOS score. Doing a statistical check like this before combining the scores from different listening levels is also recommended by the P.800 ACR procedure [5].

This added rigidness to the volume selection methodology seems to have had the intended effect of participants performing the test at the highest comfortable volume, as the rest of the participants all chose 100% volume except for one participant that chose 75%. Some would maybe argue that this could suggest that participants picked a higher volume than they were actually comfortable with, but we find this to be unlikely as the participants were clearly told that they should not pick a volume that is uncomfortable to them.

The information paper given to the participants also explains how the samples are presented, as well as showing this by including the illustration shown in Figure 3.6. This figure is translated to English, while the information paper given to the participants contains a Norwegian version. The information paper explains how each sample is comprised of two sentences spoken by the same speaker divided by 0.5 seconds of silence. This is in done in accordance to the P.800 recommendation's DCR procedure [5]. As explained in Section 3.4.2 on Audio Selection, two male and two female speakers have been selected to repeat the same four sentences. These sentences are then combined in pairs into two audio samples per speaker. Figure 3.6 shows how each of these samples are presented in each round (null-pair

Participant age: _____ **Volume:**
Participant sex: Male [ ] | Female [ ] **Participant number:**

MOS Test - **ROUND ___**

Female Speaker 1 (Speaker 1)

Female Speaker 1 (Speaker 1) – Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 1 (Speaker 1) – Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 2 (Speaker 2)

Female Speaker 2 (Speaker 2) - Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 2 (Speaker 2) - Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Page: 1/2

Figure 3.5: Page one of the two page document given on paper to the participants for each steganography method and the null-pairs.

or audio steganography method).

First, the reference audio sample is played, in the meantime the conductor of the study is holding up a sign labeled "Reference" in both English and Norwegian. Then after 1 second of silence, in accordance to the P.800 DCR procedure [5], a degraded version of the sample is played while the conductor of the experiment holds up a sign labeled "Degraded" in both English and Norwegian. By degraded sample we mean that the sample is either a null-pair, i.e. the same high quality reference, or has been modified by one of our four chosen audio steganography methods. This playback sequence is repeated twice so that the participants can "double check" if they actually noticed a degradation in audio quality between the samples. This is also done in accordance to the P.800 DCR procedure which recommends either presenting the samples in a A-B or A-B-A-B repeated sequence configuration, where A is the high quality reference and B is the degraded sample [5]. After each of these sequences the participants marks their rating on the rating paper illustrated in Figure 3.5. They know when its time for this as the conductor is no longer holding up any signs. The conductor of the study also pays attention to make sure that the rating is given for the correct sample.

A total of 40 of these sample combination files were created manually in Audacity to be played and rated during the DMOS test. This process is repeated until the test is done and the entire test takes about 30 minutes, which is within the 45 minute maximum suggested by the P.800 ACR procedure, but outside the ideal recommendation of not exceeding 20 minutes [5]. However, we think this is a worthy trade-off for having each of our participants rate all of our samples. In addition to this, the participants did not seem excessively fatigued after the test, in our subjective opinion.

While designing the experiment we noticed that two of the audio steganography methods appear to be very difficult to tell apart from the reference files, while two of them appear to be very easy: However, both the author of this thesis and two of the supervisors thought that they might have heard some differences in the files that are hard to distinguish from the reference. We therefore worry that playing the methods that are very easy to tell apart from the reference files first would ruin the sensitivity of the test, because participants would expect large deviations from the reference and therefore give the methods that are hard to distinguish a perfect score. Because of this we decide to deviate a bit from the P.800 DCR recommendation here [5]. The P.800 DCR procedure states that only one random order of presentation needs to be used in a DMOS test. However, we decide to use the same pre-defined order of presentation where all of the null-pair samples will be presented first, followed by all of the samples by the rest of the methods in order ranked from being the hardest to the easiest tell apart from the reference by us subjectively for all participants. This gave us the order of Null-pair samples, Steghide [30] samples, Hide4PGP [32] samples, GAN [37] Low samples and GAN [37] High samples.

We feel that this also helps the participants regain some focus at a crucial time, when being presented the samples withy he much more noticeable degradations,

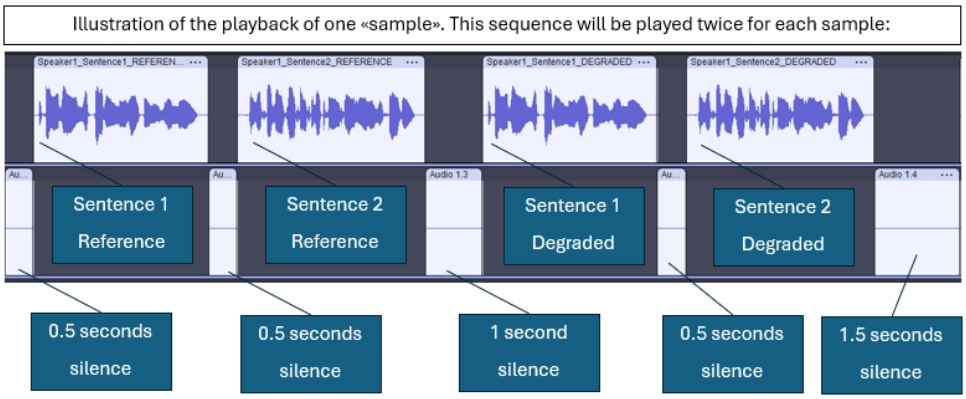


Figure 3.6: Sample presentation illustration used in the participant information paper translated to English.

and hope that this counters some of the potential listening fatigue caused by a longer than optimal test duration. It was observed while conducting the experiment that participants that appeared to be getting fatigued and tired of the test after the first three methods, that are either hard or impossible to distinguish, seemed to perk up a bit and appeared to regain some focus after hearing the large degradations of the last two methods.

The samples are manually played by the conductor of the experiment by opening different labeled audacity files. This works well as it naturally gives the participant a bit of time to rate the sample they just listened to before the next one is played. The conductor of the experiment makes sure that the participant has crossed out their rating, and that they have rated the correct sample before playing the next one. This manual method also allows for questions to be asked by the participant between samples during the experiment.

Table 3.4 shows what parts of the experiment design that followed the different procedures used from the ITU-T P.800 recommendations [5]. It also shows parts of the experiment design that were inspired by these procedures, but not following them completely. This is often because of some compromise, like using noise canceling headphones instead of potentially not being able to recruit enough participants if we have to invite them to a sound proof recording studio to conduct the experiment.

Some people may question why we state that we are doing a DMOS test, mainly following the P.800 [5] DCR procedure when Table 3.4 makes it seem like there are more experiment activities attributed to the ACR procedure than the DCR procedure. This is because the most important parts of the experiment, such as using reference and degraded samples and the degradation category rating scale, are based on the DCR procedure. As stated earlier, we try to follow the ACR procedure when something is not specified by the DCR procedure.

Procedure Followed	Part of the Experiment Design
ITU-T P.800 DCR	Degradation category rating scale formulation.
ITU-T P.800 DCR	Sample & reference presented in A-B-A-B format.
ITU-T P.800 DCR	Silence duration between samples.
ITU-T P.800 DCR	Samples consists of two sentences from same speaker.
ITU-T P.800 DCR	Speech material from 4 speakers.
Inspired by DCR	Order of presentation of samples.
ITU-T P.800 ACR	About 2-3 second long sentences.
ITU-T P.800 ACR	Studio quality, easy to understand speech material.
ITU-T P.800 ACR	Separated male and female MOS scores.
ITU-T P.800 ACR	Separated male and female MOS scores.
ITU-T P.800 ACR	Non-technical instructions to subjects.
ITU-T P.800 ACR	Participant requirements.
ITU-T P.800 ACR	Below 45 min maximum test duration.
Inspired by ACR	Noise-canceling headphones.
Inspired by ACR	Listening environment.
Inspired by ACR	Volume selection method.

Table 3.4: The different ITU-T P.800 [5] procedures followed and used as inspiration for the different experiment activities of the Subjective DMOS test.

3.6.2 Experiment Setup and Technical Details

To ensure consistency throughout the experiment, the same laptop is used across its entire conduction. This laptop is a Lenovo ThinkPad L13 Gen 2, using the onboard Realtek soundcard for 3.5 mm output set to its maximum output quality of 2 channel 24-bit 48000 hz to power the headphones used in the test. The laptop is running Windows 11 Pro and the audio samples are played using Audacity [73]. Furthermore, the participants listen to the audio samples using a pair of Sony WH-1000XM3 Noise canceling headphones with noise canceling turned on. The headphones are connected to the laptop via a 3.5 mm cable, a 3.5 mm splitter is used to split the output signal to a 3.5 mm Eletra CA101 chat headset [79] that is worn by the conductor of the experiment. The Eletra headphones have a quite low frequency rate at 32 ohms, but it is still observed that using the two headphones together lowers the volume of both of them noticeably.

The extra headphones are used so that the conductor can monitor the playback and hear if there are any disturbances caused by other factors that the tested degradations. The Eletra headset's volume adjustment wheel is taped to its maximum position to ensure consistency throughout the experiment. As mentioned in Section 3.3 on practical considerations for the experiment, disposable non-woven sanitary headphone covers are used on the ear-pads of the Sony headphones for all participants. It is possible that this could affect how the audio sounds both by changing the isolation capabilities of the ear-cups, and by covering some of the noise-canceling microphones. However, we believe that the microphones are still

largely able to do their job as it is quite normal to use wind-muffs and pop-filters on microphones. All experiments are done at a table or desk where the participant is seated in a way that they can not see the computer screen of the laptop used by the conductor. This is done to prevent them from realizing any details about the test that they should not, such as the first round of samples being a null-pair test.

3.7 Objective MOS Testing Methodology

3.7.1 Selection of MOS-LQO Algorithms

This [2] review of the state of the art withing the field of audio steganography is used to find the most commonly used MOS-LQO algorithms used for evaluating perceptual transparency within the field. This survey reviews 134 audio steganography papers, out of these 134 papers 12 of them use PESQ [18] as an evaluation metric. The only other MOS-LQO algorithm mentioned in this survey is PEAQ [10], but only three of the articles surveyed used this in their evaluation.

During the pre-project for this thesis [9], we also discovered some other MOS-LQO algorithms that could be used for this purpose. These were, POLQA [80], ViSQOL [12], and AqUA [13]. As mentioned in the Background Chapter in Section 2.6, POLQA and AqUA is excluded from our study because we were not able to acquire academic licenses for them, and PEAQ [10] is excluded as we are unable to find its license conditions. Section 2.6 also mentions how PESQ can be used in this study, despite being a paid piece of software, because of an exception in its license, and that ViSQOL Speech [12] and ViSQOL Audio [48] are both open source and free to use for anyone. More details about these MOS-LQO algorithms, and why some of them were excluded from this study, can be found in Section 2.6.

This leaves us with PESQ, ViSQOL Speech and ViSQOL Audio as the selected MOS-LQO algorithms to be evaluated on their ability to measure the perceptual transparency of audio steganography methods in this thesis.

3.7.2 PESQ Implementation

A PESQ wrapper for Python by ludlows on GitHub [46] is used as our PESQ implementation. This contains a modified version of the PESQ original source code, and a wrapper to be able to use it in Python. This is a popular way to use PESQ and is used by several big actors such as NVIDIA, Facebook Research and SpeechBrain [46]. ludlows implementation supports both wide band and narrow band PESQ, we are only using wide band PESQ in this thesis as we are working exclusively with sample rates of 16.000 kHz or higher.

Wide-band PESQ requires 16.000 kHz sample rates in order to function properly [46]. The same 16.000 kHz cover files were therefore used for all comparisons. The stego files generated by the GAN steganography method have to be resampled from 44.100 kHz to 16.000 kHz. We also convert them back to 16-bit PCM files as we had problems with their 32-bit float format while measuring SNR

and figure 16-bit PCM files appear to be more common and less likely to cause problems that we may or may not be able to easily notice. Despite these conversions the author of this thesis can not tell the difference between the converted and non-converted audio samples when listening to them side by side, so we assume that the effect this conversion has on the PESQ measurement is minimal. The resampling also has to be done to measure PESQ correctly at all, so this is an easy decision to make.

A Python Jupyter notebook is used to apply the Python wrapper [46] in order to measure the PESQ scores of all of our stego files. Microsoft CoPilot [20] is used to help set up the requirements for the wrapper, and to generate code applying PESQ to our stego files and writing the calculated scores for each stego file to CSV files for each of our steganography methods. The wrapper requires a C compiler to be installed and to be accessible in Visual Studio Code (VSCode) [81], as this is the program we used for developing our notebook. "MSCV Compiler", "Windows 11 SDK" and "CMake" are downloaded through Microsoft's C++ build tools installer [82], as prompted by CoPilot, to full-fill this requirement. Then the conda environment for the project used to run the Python Jupyter notebook is activated in a terminal, and the following commands are ran. First:

```
"C:\Program Files (x86)\Microsoft Visual Studio\2022
\BuildTools\VC\Auxiliary\Build\vcvarsall.bat" x64
```

Is ran to to activate the C compiler in the environment. Then:

```
code .
```

Is ran to launch VSCode directly from the terminal.

After doing this the Python Jupyter notebook is executed and the CSV files containing the scores for all of our stego files are generated. The Python Jupyter notebook containing the code used for this is available on our GitHub [67] in the "PESQ Code" directory.

3.7.3 ViSQOL Implementation

Google ViSQOL [48] contains two modes that are both used in this thesis; ViSQOL Speech and ViSQOL Audio. ViSQOL Speech requires the audio files to be evaluated to have a sample rate of 16.000 kHz, while ViSQOL Audio requires 48.000 kHz sample rates to be used. A set of 16.000 kHz and 48.000 kHz resampled cover files in 16-bit PCM WAV format are therefore created with Audacity [73] to be used as reference files for all stego files from our different steganography methods. Resampled stego files with 16.000 kHz and 48.000 kHz sample rates from all of our methods are naturally also created with Audacity [73] in order to be able to evaluate them with the ViSQOL MOS-LQO algorithms.

Google ViSQOL version 3.3.3 is downloaded from the official GitHub repository [47], and is ran on an Ubuntu 22.04 LTS WSL 2 [72] virtual machine (VM).

The ViSQOL repo's guide for installation is followed to set up the software. The prerequisites Numpy [83] and Bazel [84] are installed and ViSQOL v 3.3.3 is built from its source code using Bazel. Bazel initially returns a 404 error for fetching a file called "armadillo" while trying to build ViSQOL, stating that it cannot be found. Microsoft CoPilot [20] is used to help with debugging and points us to this GitHub issue [85] where the answer from user "rsanchezpizani" is followed to resolve the issue. This answer states to download another version of armadillo from the website [86] (version 14.0.3), then get the sha256 checksum for this new version and update the checksum, link to the download and version name in the WORKSPACE file used by Bazel to build the software. Applying this fix resolves the issue and allows us to successfully build ViSQOL.

Google ViSQOL's GitHub repository (repo) shows some example commands where a batch input CSV file is used. This seems suitable for our purposes, but we could not find information on how to structure the CSV file. Microsoft CoPilot is again consulted to resolve this issue and replies that the CSV batch files should contain two columns with the top rows containing the words "reference" and "degraded", with file paths for the reference and degraded file pairs following in the columns below. It also stresses the importance of LF line endings being used for our Linux VM instead of the CRLF line endings that are used for Windows, and shows us how to change these line endings in VSCode by pressing a button at the bottom of the screen. Eight CSV batch files are created to specify the paths of both the 48.000 kHz and 16.000 kHz audio samples from all of our steganography methods to be evaluated by ViSQOL Speech and Audio. The evaluations are then ran using these commands as specified by the guide in the ViSQOL GitHub repo [47]. For ViSQOL Speech this command is ran for each of our stego method's ViSQOL Speech CSV batch files:

```
./bazel-bin/visqol --batch-input_csv [BATCH CSV PATH]
--results_csv [PATH FOR RESULTS] --output_debug [DEBUG OUPUT PATH]
--use_speech_mode
```

And for ViSQOL Audio the following command is ran for each of our stego method's ViSQOL Audio CSV batch files:

```
./bazel-bin/visqol --batch-input_csv [BATCH CSV PATH]
--results_csv [PATH FOR RESULTS] --output_debug [DEBUG OUPUT PATH]
```

This leaves us with eight CSV files containing the ViSQOL Speech and ViSQOL Audio results of the evaluations of all of the audio samples for all our four tested audio steganography methods.

3.8 SNR Implementation

Our Signal to noise ratio (SNR) implementation is almost entirely based on the "compute_transparency.py" code [87] used Reyer's thesis [3]. The only changes

made was the changing of some paths and repetition of some parts of the code to generate one CSV file for each of our stego methods, containing the SNR scores for all of their respective stego files.

Reyers SNR method [87] works by first splitting up the cover and stego audio files into the (in our case 16-bit) samples they are composed of. Then it computes the Mean Square Error (MSE) of the difference between the two sets of samples, before finally dividing the cover file samples with the MSE and multiplying the result by $10log_{10}$ to compute the SNR in decibel. Reyers defines their SNR computation like this in their thesis [3]:

$$SNR = 10\log_{10}(\frac{XS}{MSE}),$$

$$\text{where } MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

$$XS = \frac{1}{N}\sum_{i=1}^{n}x_i^2$$

"Where x and y are the cover and stego audio signals respectively, N is the number of samples in a signal, and x_i and y_i are the i^{th} sample of x and y respectively." [3, p. 6].

After running the SNR calculations we are left with one CSV file for each of our stego methods which contains the SNR results of all stego files for each method. The average of these scores are also taken manually in Microsoft Excel to be used in Results Chapter (Chapter 4).

The exact code we used for our SNR implementation is available on our GitHub [67] in the "SNR Code" directory.

3.9 Method for Calculating DMOS Scores

This section explains the methodology we have used to calculate our Subjective DMOS scores, as well as all the statistical methods we applied to our data. Microsoft CoPilot was used to discuss different options, but all chosen methods were discussed with supervisors and cross checked with other sources. The ITU-T P.800 recommendations on calculating mean opinion scores were naturally also used [5]. For the statistical tests requiring a confidence interval, we decide to use a 95% interval, as this appears to be commonly used in computer science [88]. This means that our significance level (α) for our P-value is set to 0.05 [88]. The P-value stands for probability value, and tells us the strength of the evidence against our null-hypothesis [89]. The P-value says something about the probability of obtaining equally or more extreme results than what is observed, if one assumes that the null hypothesis is correct [89].

In accordance with the ITU-T P.800 recommendations for MOS testing [5], DMOS scores are calculated by simply taking the average of each participants ratings of all samples for each degradation, which in our case are our audio steganography methods. DMOS scores can also be calculated separately for male and female samples by just taking the averages of these respective samples separately [5]. For the sake of identifying outliers we will also be using the intermediate DMOS scores from each participant for each sample to use with a statistical method that is often used for this, this will be further explained in Section 3.9.1 following below.

3.9.1 Excluding Outliers

The temporary DMOS scores from all the participants are first calculated in Microsoft Excel following the methodology explained here: First, all participants ratings for all male and female audio samples for the null-pairs and our four steganography methods are used to calculate separate male and female sample DMOS scores for each steganography method and the null-pairs. This is done by taking the average of all scores for all female samples and male samples for each participant to get participant isolated DMOS scores for male and female samples separately.

These individual DMOS scores are used to calculate Z-scores used to identify potential outliers. Z-scores let us know how many standard deviations above or below the average a specific DMOS score is [90], a common way to identify outliers is to set an absolute Z-score value threshold for extreme values at either 2 or 3 [90]. We set our threshold at 2 because we had subjective observations to back up this choice, which we will get further into later in this section. The Z-scores are calculated following this equation, and this tutorial [91] is used to calculate it in excel:

$$Zscore = \frac{IndividualDMOS - MeanDMOS}{StandardDeviation}$$

Z-scores are calculated for all participant's male and female sample DMOS scores and two potential outliers are discovered. The first potential outlier gets z-scores of -2.83 and -3.46 for null-pair male and female DMOS scores respectively, as well as -3.45 and -2.53 for male and female Steghide null-pair scores respectively, and -2.92 for their female Hide4PGP DMOS score. This means that they are consistently ranking these samples to be way lower quality than the average participant. Because of these extreme discrepancies, and especially since the null-pair samples are rated so much lower than by the average participant, we decide that we cannot trust this participants ability to rate the audio samples consistently, and decide to exclude all of their scores from the test entirely.

The second potential outlier gets less extreme Z-scores for the null-pair, Steghide and Hide4PGP DMOS scores. However, they get Z-scores of 1.97 and 2.05 for male and female GAN Low DMOS scores respectively, and 2.06 and 2.08 for male and female GAN High DMOS scores respectively. This is close to the threshold,

and it would be possible to set the Z-score threshold at 3 to keep this participant. However, this participant also picked a volume level of 12%, where all other participants, except one that picked 75%, picked 100% volume. This low volume participant also expressed that they could barely hear any degradation in the GAN audio samples, and wanted to hear the samples again after the test with higher volume. Upon doing this the participant stated that they would have rated the samples way worse if they had listened at a higher volume.

In addition to this, the low volume participant's average DMOS scores for both GAN Low and GAN High was 5 for female samples and 4.75 for male samples. Where the average DMOS scores of all participants (before excluding outliers) was 2.58 and 2.69 for female and male GAN Low samples respectively, and 2.5 and 2.77 for female and male GAN High samples. If we look back at the DCR rating scale used for DMOS in Figure 3.4 we can see that the second potential outlier participant rated these methods as having either inaudible or close to inaudible degradations, while the average participant rated them as being somewhere between annoying and slightly annoying.

Judging from these qualitative observations, we can clearly tell that a Z-score threshold absolute value of 2 is not excessive in our situation. The combination of all of these observed discrepancies makes us convinced that we can not trust this second potential outlier participant's ability to hear degradations at the volume they were using during the test. This leads us to exclude also this participant from the test entirely.

After excluding these two outlier participants from the test entirely, the individual DMOS scores are recalculated. Male and female average DMOS scores across all non-excluded participants are calculated as the actual Male and Female DMOS scores, and the average of these as the combined DMOS score, which is usually the one just referred to as the DMOS score.

The Excel sheet used for our Z-test calculations can be found on our GitHub [67] in the "Supplementary Materials" directory.

3.9.2 Statistical Difference Between Male and Female Sample DMOS Scores

The ITU-T P.800 recommendation states that male and female ACR MOS scores can only be combined if they are not significantly different. While the P.800 DCR procedure does not explicitly mention this, it also does not explicitly say not to, and we can not see a reason why the DMOS scores could be combined, and not the MOS scores. We therefore decide to follow the P.800 ACR procedure's recommendations, and check if our male and female sample DMOS scores are statistically different to see if we can report our results using just the combined DMOS score.

Microsoft CoPilot [20] is consulted for suggestions of tests to use to assess if a statistically significant difference is present, and the suggestions given are cross checked with other sources. A Students T-Test is initially suggested, and this You-

Tube tutorial [92] is used for extra information and to learn how to perform one in Microsoft Excel. The tutorial states that we would have to use a paired t-test since we are testing if two conditions are rated differently by the same population. [93] confirms this, stating that paired difference tests should be used if we are working on paired data, and use experiments where the same participants are used to compare two different conditions. [93] mentions the paired-samples t-test and Wilcoxon signed-rank test among others as some examples of paired difference tests that can be used on paired data like this.

According to [94] and [95, p. 344] a paired t-test assumes that the the differences in measurements are at least approximately normally distributed. We therefore need to see if this is true for our data before proceeding with the students T-test. To check if this is the case, CoPilot suggests we perform a Shapiro-Wilk test, and [96] confirms that this is a commonly used statistical method for checking normality for datasets with less than 5,000 samples.

We do our Shapiro-Wilk test by plotting the differences between male and female samples for all of our measured methods and the null-pairs into this [97] online calculator and setting a significance level of 0.05, to align with our 95% confidence interval defined in Section 3.9. This gives us P-values of 0.0087, 0.0015, 0.0156, 0.0364 and 0.5199, for Null-pair, Steghide, Hide4PGP, GAN Low and GAN High respectively. The GAN High scores are above the threshold and are clearly normally distributed, so we can definitively do a T-test on this method, but the rest are all below the significance level. This means that these methods likely do not have normally distributed differences and that we cannot use the paired t-test for these methods. The t-test is therefore done only on the GAN Low method in Microsoft Excel following this [92] tutorial. This gives us a P-value of 0.0506 which is barely not significant if we follow a strict cutoff threshold of 0.05, which appears to be the norm for T-tests [98].

Microsoft CoPilot [20] suggests using Wilcoxon Signed-Rank tests for the methods without normally distributed differences, and this [99] online calculator is identified. [99] confirms that this test appears to be suitable for our purposes, by stating that the Wilcoxon Signed-Rank test works well on data from experiments where the samples are correlated, for investigating the difference between conditions or treatments. [99] also gives an example of a suitable experiment where some data scoring reading performance from the same children before and after reading training is evaluated to be significantly different or not. This Norwegian university level statistics book also confirms that a paired Wilcoxon test is a good way of assessing if of two paired datasets are significantly different [95, pp. 360–362].

One problem we found with the Wilcoxon Signed-Rank test is that, according to [99], there should optimally be no ties in the data to ensure maximum accuracy, while we have quite a few ties in our differences. Ties, meaning data points with the same differences so that their ranks are tied in the test [100]. However, we still chose to use this method as it does have ways to adjust for ties [99, 100]. The biggest problem is tied differences of 0, as these are typically omitted from the

test [101]. This can cause issues because the number of remaining values need to be large enough for the method's "Wilcoxon W statistic" to create a normal distribution in order to be able to use the results [99]. The null-pair and Steghide samples fail to generate usable results for this reason. However, Hide4PGP gets a P-value of 0.00466 which is far below the 0.05 threshold. Since this is so far below the threshold we think it shows that the male and female sample DMOS ratings for the Hide4PGP method are very likely significantly different, even with the potentially lower accuracy caused by the ties. We also think that it is better to be wary of combining the DMOS scores, rather than to potentially lose accuracy by going against the ITU-T P.800 recommendations [5] by combining the scores without definitively proving that they are not statistically different, and potentially losing accuracy in our results.

To further confirm a significant difference between male and female samples for our Hide4PGP results, we also perform a paired sign test, which according to [95, p. 362] requires paired data, but makes no other assumptions about it, coming at the cost of sensitivity. A paired sign test is done on all of our method's male and female sample DMOS scores from all participants using this [102] online calculator. P-values of 1, 0.4795, 0.00228, 0.2059 and 0.0707 are achieved for null-pairs, Steghide, Hide4PGP, GAN Low and GAN High respectively. This very low P-value of 0.00228 is far below our 0.05 significance threshold, and again strongly confirms the Wilcoxon signed rank tests strong indication of there being a significant difference between male and female sample DMOS scores for the Hide4PGP audio samples.

Since the individual male and female DMOS scores across all our participants for one of our steganography methods are shown to very likely be statistically different, we will report all DMOS scores generated by all methods by gender to keep things consistent. We will refer to these as $DMOS_M$ and $DMOS_F$ for male and female DMOS scores respectively. We will also often report the concatenated DMOS scores as just $DMOS$, along with the separated ones.

3.10 Methods for Comparing Results

3.10.1 Pearson's Correlation and Mean Absolute Errors

After discussing pros and cons of different methods and metrics for quantifying the results of this thesis in a meaningful way with Microsoft CoPilot [20], and looking at other sources [103, 104] for more accurate information about the suggested methods, Pearson's correlation [103] and mean absolute errors (MAE) [104] were chosen to be used for this purpose.

The statistics programming language R [105] is used in the RStudio IDE [106] with code largely generated by Microsoft CoPilot [20] to calculate Pearson's correlation, after using this [107] YouTube tutorial to understand the basics of Pearson's correlation in R. Microsoft Excel is used to calculate the MAE values following this [108] YouTube tutorial.

Mean Absolute Errors or MAE is calculated just how it sounds. The average absolute differences between the DMOS scores and each of the MOS-LQO algorithm scores for all audio samples are calculated for each of our stego methods. We also calculate separate MAE scores for male samples and female samples for each method in the same way, by only including the relevant samples in the calculation. This leaves us with three MAE scores for each MOS-LQO algorithms, which show us the absolute differences between DMOS and the scores from the algorithms. In addition to MAE, manual inspection will be done to see if the errors go largely in one direction (negative or positive), as this also might be relevant for evaluating the MOS-LQO Algorithms. The Excel sheet used for our MAE calculations can be found on our GitHub [67] in the "Supplementary Materials" directory.

Pearson's correlation is taken between the $DMOS$ scores for all 32 audio samples and the MOS-LQO scores for the sample samples, and for the 16 $DMOS_{Sex}$ scores and $MOS-LQO_{Sex}$ scores each for the male and female samples. Giving us a total of nine correlation scores showing how closely our three tested MOS-LQO algorithms PESQ, ViSQOL Speech and ViSQOL Audio correlate with DMOS overall for all of our samples, and for male and female audio samples separately. The confidence level of the results is also calculated in R, and scatter plots with regression lines are created for each of the nine results.

The Pearson's correlation values, 95% confidence intervals and p-values are produced using the built in cor.test function of R like this:

```
result1 <- cor.test(df$DMOS, df$PESQ, method = "pearson")
```

This returns an output like this:

```
Pearson's product-moment correlation

data:  df$DMOS and df$ViSQOL.A
t = 26.977, df = 30, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9590322 0.9902935
sample estimates:
      cor
0.9800042
```

The confidence level value to be put besides the score is then calculated with R like this:

```
margin_of_error1 <- (result1$conf.int[2] - result1$conf.int[1]) / 2
formatted_result1 <- paste(round(result1$estimate, 4), "±",
 round(margin_of_error1, 4))
```

And is returned along with the Pearson's correlations score (r) like this:

```
> formatted_result1
[1] "0.9649 ± 0.0271"
```

Which is also the value we report in our Pearson's correlation results in Section 4.3.1. As stated previously, the R-code for calculating Pearson's correlation is largely generated by Microsoft CoPilot [20].

The exact R-code used can be found on our GitHub [67] in the "R Correlation Code" directory.

3.10.2 Method for analyzing SNR results

A qualitative analysis is done, comparing the SNR and DMOS scores on a audio sample level, to see if the identified 30 dB threshold from [8] and [3] holds up. The SNR scores of each audio sample are manually compared to their respective DMOS scores, paying particular attention to the samples that scored poorly on the DMOS test. We then use these comparisons to assess how likely we deem the claimed 30 dB threshold for human perception found in the literature to be correct.

We also pay particular attention to the DMOS rating scale given to our human participants, that can be seen in Figure 3.4. By doing this we can see if the degradations of a sample that has achieved an SNR score above 30 dB, since higher SNR scores typically translate to better perceptual transparency [3], is likely to be perceivable by our human participants. If the participants for instance rank the degradation of a sample as "Annoying", according to the DMOS rating scale, and this sample has an SNR score above 30 dB, we can be pretty certain that the identified threshold does not hold true in all situations.

Chapter 4

Results

This chapter explains the results from our different completed research activities done to answer our research questions. The chapter starts by outlining the DMOS results, before moving on to the results measured by our MOS-LQO algorithms. It then compares the results from all of our three chosen MOS-LQO algorithms and DMOS separately. It does this by presenting and discussing patterns in our measured Pearson's correlation results, mean absolute error (MAE) results and our various other results. It then goes on to present and discuss the SNR scores, and compares these scores to our DMOS scores to assess whether the identified threshold of 30 dB discovered in previous work [3, 8] holds up. Finally, a mathematical error found in the paper [19] proposing our implemented TAN audio steganography method is presented.

4.1 DMOS Results

4.1.1 Participant Demographics

Even though it is not required by the ITU-T P.800 [5] or the P.800.1 [6] recommendations for MOS testing, it seems to be considered best practice in some fields, such as using MOS testing to determine the naturalness of synthetic speech [58], to publish listener demographics. We also thought this could be useful information in the field of audio steganography as many different things can affect a subjective study, which the different elements that can affect speech quality illustrated in Figure 3.1 is a good example of. We therefore report the listener demographics of our DMOS test in this Section, excluding the outliers identified in Section 3.9.1.

All 19 listeners are proficient in English, and most of the participants are native Norwegian speakers. No participants are native English speakers, but the exact native languages of the participants are omitted for privacy reasons. All audio samples used in the DCR DMOS test are spoken in English and, as mentioned in Section 3.4.2, and audio samples that are easy to understand are selected, which may address concerns that could arise from non-native English speakers being used as listeners to some degree. Since the listeners are instructed to rate the

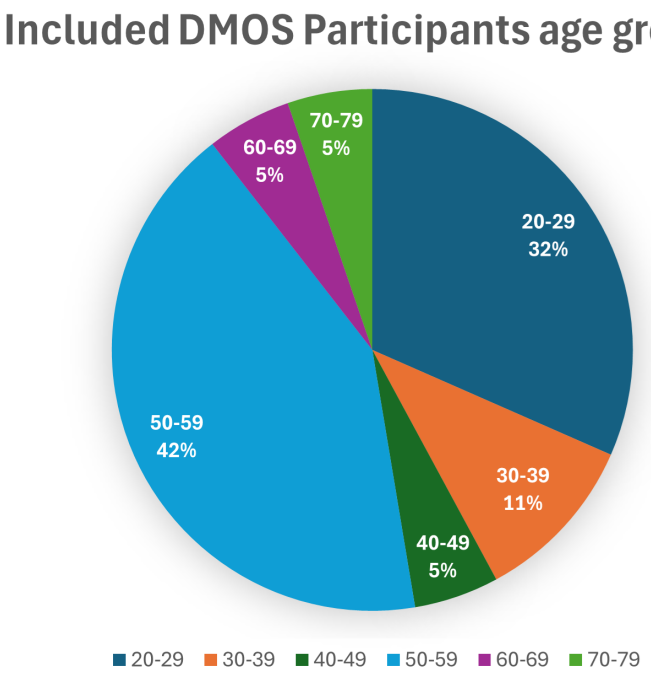


Figure 4.1: The subjective DCR DMOS test participants age demographics.

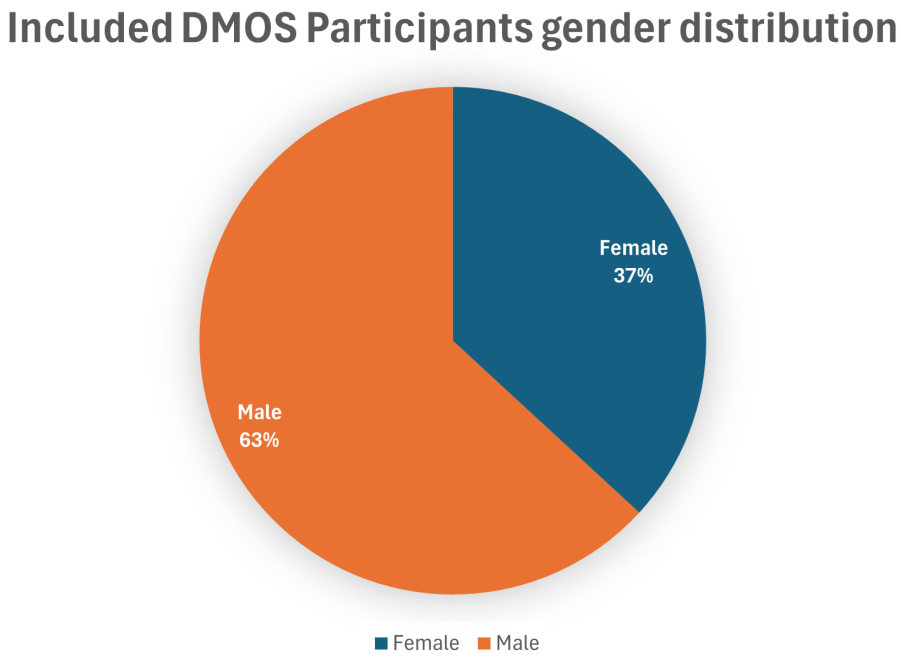


Figure 4.2: The subjective DCR DMOS test participants gender distribution.

degradation in audio quality between a high quality reference and a degraded sample, we think that the listeners native languages probably affect the results less than if they for example were to rate something like speech quality or naturalness.

The age distribution of the participants of the subjective DCR DMOS test are reported in a way that respects privacy, by creating age groups with intervals of ten years from the ages 20 to 79. The age demographics are displayed in the pie chart in Figure 4.1. The average rounded off age is 42, and the median age is 50. The listeners gender distribution is also shown in a pie chart in Figure 4.2, showing 63% male and 37% female participants.

4.1.2 DMOS Scores

As mentioned in Section 3.9.2, DMOS scores for female and male samples are reported along with the concatenated DMOS score of all samples. These scores are referred to as $DMOS_F$, $DMOS_M$ and $DMOS$ respectively. As explained in Section 3.9, DMOS scores are calculated by taking the average of the ratings of all audio samples, from every one of our participants, for each degradation separately, to calculate the overall $DMOS$ scores for each degradation. The same thing is done to the male and female samples separately to calculate the $DMOS_M$ and $DMOS_F$ scores. The results of these calculations for each degradation, degradations being the steganography methods and null-pair test, are displayed in Table 4.1.

Degradation	$DMOS_F$	$DMOS_M$	$DMOS$
Null-Pair	4.7895	4.7895	4.7895
Steghide	4.7105	4.7895	4.7500
Hide4PGP	4.5921	4.8158	4.7040
GAN Low	2.5658	2.6579	2.6118
GAN High	2.4605	2.7500	2.6053

Table 4.1: The DMOS results from the subjective DCR DMOS test.

A box plot illustrating the distribution of $DMOS$ scores across all male and female samples separately, for each stego method and the null-pairs, can be seen in Figure 4.3. Simply put, the boxes of the box plot shows where most of the given DMOS scores lie, the whiskers show values that deviate from this without being outliers, the lines through the boxes show the median scores, and the red dots show outlier scores [109] identified by our Z-test in Section 3.9.1. The box plot was created with this [110] online tool, and the red outlier ratings were added manually with this online photo editor [111]. The outlier ratings are taken from the specific ratings from the two excluded participants that were identified to be too far outside the norm by the Z-test in Section 3.9.1. Since these participants were completely excluded from the study, their other ratings are not included in the box plot. Their outlier scores are simply shown on the box plot to illustrate how far they deviate from the norm, and to further justify their exclusion visually.

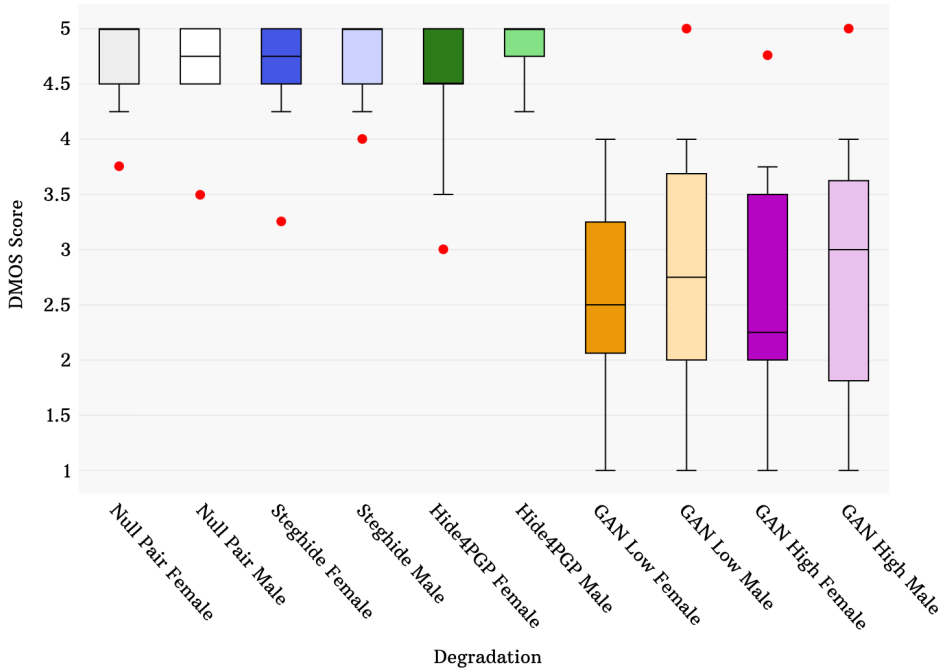


Figure 4.3: Box plot illustrating the distribution of DMOS scores given by the participants of our DCR DMOS experiment. DMOS scores are shown on the Y-axis, and degradation samples (stego methods and null pairs) by gender on the X-axis.

By inspecting the box plot in Figure 4.3 we can clearly see how the Hide4PGP results were quite varied across the samples our from male and female speakers, while the other methods and null pair vary to a lesser degree. This adds more credibility to the results found in Section 3.9.2 of this thesis, showing that the $DMOS_F$ and $DMOS_M$ results are significantly different for the Hide4PGP method only. It also further shows that reporting $DMOS$, $DMOS_F$, and $DMOS_M$ scores separately, rather than just reporting the concatenated $DMOS$ scores is likely beneficial for the accuracy of our results. More about how we used statistics to come to this conclusion initially can be read in Section 3.9.2.

The raw DMOS test results and score calculations can be found on our GitHub in the "Supplementary Materials" directory [67].

4.2 MOS-LQO Algorithm Results

The MOS-LQO Algorithm results refers to the MOS-LQO scores generated by our chosen MOS-LQO algorithms; PESQ [18], ViSQOL Speech [48], and ViSQOL Audio [48]. Since we want to compare these algorithms to the DMOS scores, using the DMOS scores as sort of a ground truth for how these algorithms should per-

form, as justified by [12], we will also report male and female sample scores separately for these algorithms, as well as a concatenated score with the average of both genders. The MOS-LQO scores are calculated similarly to the DMOS scores by taking the average of the scores produced for each sample by each algorithm for both male, female and all samples separately. We will label these MOS-LQO scores as the name of the algorithm marked by either M, F or nothing to indicate if its the score for male only, female only or all samples. For PESQ for instance this would like like this: $PESQ_F$, $PESQ_M$, $PESQ$. The results for PESQ are displayed in Table 4.2, the results for ViSQOL Speech are displayed in Table 4.3, and the results for ViSQOL Audio are displayed in Table 4.4. The raw results from the MOS-LQO algorithms for each sample can be found on our GitHub [67] in the "Supplementary Materials" directory.

Degradation	$PESQ_F$	$PESQ_M$	$PESQ$
Steghide	4.6257	4.6323	4.6290
Hide4PGP	4.6405	4.6391	4.6398
GAN Low	3.9602	3.9690	3.9646
GAN High	3.9580	3.9631	3.9606

Table 4.2: The PESQ results from the objective MOS-LQO testing.

Degradation	$ViSQOLSpeech_F$	$ViSQOLSpeech_M$	$ViSQOLSpeech$
Steghide	4.4697	4.5143	4.4920
Hide4PGP	4.4987	4.5206	4.5097
GAN Low	3.5624	4.1495	3.8560
GAN High	3.5455	4.1387	3.8421

Table 4.3: The ViSQOL Speech results from the objective MOS-LQO testing.

Degradation	$ViSQOLAudio_F$	$ViSQOLAudio_M$	$ViSQOLAudio$
Steghide	4.7218	4.7275	4.7247
Hide4PGP	4.7305	4.7282	4.7293
GAN Low	1.9362	2.0022	1.9692
GAN High	1.9450	1.9984	1.9717

Table 4.4: The ViSQOL Audio results from the objective MOS-LQO testing.

4.3 DMOS vs MOS-LQO Results

In this section we will compare the DMOS and MOS-LQO results through different means to see how well the MOS-LQO algorithms align with the subjective DMOS

results. We do this comparison, as subjective mean opinion score results are typically considered the ground truth for MOS testing [12], and our tested MOS-LQO algorithms appear to have been made to emulate subjective P.800 [5] ACR MOS tests [12, 15], which share a lot of similarities with P.800 DCR DMOS, as we exaplained in Section 3.2. Our full reasoning for using DCR DMOS instead of ACR MOS can also be seen in Section 3.2, but quickly said it is mostly because of DCR DMOS's increased sensitivity over ACR MOS [5].

4.3.1 Pearson's Correlation Results

This section presents the Pearson's Correlation results. As mentioned in Section 3.10.1, the Pearson's correlation is calculated by using the R programming language. Nine scores are generated for the correlation between DMOS and our three MOS-LQO algorithm scores for; all samples, female samples and male samples respectively. The numerical Pearson's correlation results are presented in Table 4.5, along with their respective confidence levels and P-values. Scatter plots with regression lines like the one seen in in Figure 4.4 are also created with R to show the spread and correlation of the results visually. More information about how this is done can be found in Section 3.10.1. Figure 4.4 shows the scatter plots generated for all samples rated by the MOS-LQO algorithms, larger versions of these plots, as well as isolated plots for male and female samples, can be found in Appendix G.

Scores compared	Correlation and confidence level	P-value
$DMOS\text{-}PESQ$	0.9649 ± 0.0271	$< 2.2e-16$
$DMOS\text{-}ViSQOLSpeech$	0.772 ± 0.1519	$2.284e-07$
$DMOS\text{-}ViSQOLAudio$	0.98 ± 0.0156	$< 2.2e-16$
$DMOS_F\text{-}PESQ_F$	0.9555 ± 0.0556	$8.204e-09$
$DMOS_F\text{-}ViSQOLSpeech_F$	0.8867 ± 0.1313	$4.738e-06$
$DMOS_F\text{-}ViSQOLAudio_F$	0.9899 ± 0.0132	$2.852e-13$
$DMOS_M\text{-}PESQ_M$	0.9807 ± 0.0248	$2.503e-11$
$DMOS_M\text{-}ViSQOLSpeech_M$	0.7656 ± 0.2397	0.0005465
$DMOS_M\text{-}ViSQOLAudio_M$	0.9745 ± 0.0326	$1.759e-10$

Table 4.5: The Pearson's correlation values between DMOS and our tested MOS-LQO algorithms along with their confidence levels and P-values.

As we can see from the results in Table 4.5 ViSQOL Audio clearly correlates the closest to our subjective DMOS scores across all of our results, with PESQ coming in closely behind, and ViSQOL Speech falling behind the two others by a significant amount. In addition to this, ViSQOL Audio appears to have a slightly more consistent correlation across male and female samples than the two others. We can also see by the P-values that all the MOS-LQO algorithms had statistically significant correlations to DMOS for all samples, male samples, and female

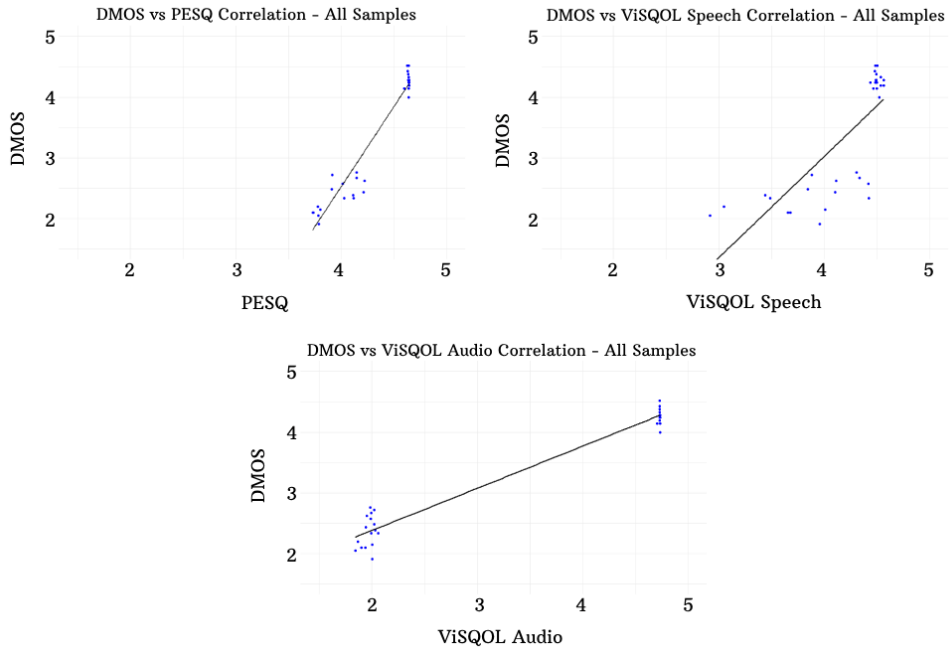


Figure 4.4: These scatter plots show the overall distribution and correlation between the ratings generated by our three MOS-LQO algorithms and DMOS for all samples.

samples when setting the significance threshold P-value at 0.05. All P-values are extremely small, strongly suggesting that the correlation results were not caused by random chance.

By inspecting the scatter plots in Figure 4.4 we can also see that ViSQOL Audio appears to consistently score overall stricter than DMOS, while PESQ and ViSQOL Speech score consistently less strict overall. This is also an interesting result, as it may go even more in favor of ViSQOL Audio, as we would probably want deviations to be stricter rather than less strict in a security conscious field like audio steganography.

4.3.2 Mean Absolute Error Results and Manual Observations

This section showcases our mean absolute error (MAE) results between DMOS and our tested MOS-LQO algorithms; PESQ, ViSQOL Speech and ViSQOL Audio. As mentioned in Section 3.10.1 the MAE between our DMOS and MOS-LQO scores show how much the MOS-LQO scores deviate from the DMOS scores on average. MAE uses absolute values, so we can therefore not see the direction of the deviation, but rather all deviations quantified into a single positive metric. We will therefore be using some manual observations of the MOS-LQO and DMOS scores from Tables 4.1, 4.2, 4.3 and 4.4 to assess if there is a overwhelming general direction of these errors.

When looking at the MAE scores of each MOS-LQO algorithm by itself in Tables 4.6, 4.7 and 4.8, we can see that ViSQOL Audio deviates on average by far the least from DMOS, this is true for all samples, all female samples and all male samples separately.

When looking at the DMOS and MOS-LQO results in Tables 4.1, 4.2, 4.3 and 4.4, we can also see that there does appear to be overwhelming general directions for all of the MOS-LQO algorithms when it comes to the low quality degradations done by the GAN methods. PESQ and ViSQOL Speech appears to consistently rank samples way less strict than DMOS for these methods. While the subjective DMOS scores suggests that these methods on average rank somewhere between slightly annoying and annoying when looking at the DMOS DCR scale in Figure 3.4, PESQ and ViSQOL Speech both rank these degradations closer to "Perceptible but not annoying" on average. ViSQOL Audio on the other hand, ranks these degradations moderately stricter than DMOS on average, rating them both ever so slightly below "Annoying" on the DCR scale, however it is still about three times closer to the DMOS rating than the other MOS-LQO algorithms.

For the degradations produced by the other stego methods; Steghide and Hide4PGP, which if we compare the DMOS results to the null-pairs pretty much have perfect scores, PESQ and ViSQOL Speech both give slightly stricter scores than DMOS on average across all samples, while ViSQOL Audio gets very close to the subjective DMOS scores, being just ever so slightly stricter.

As we can see by these results, ViSQOL Audio appears to align much closer with DMOS than PESQ and ViSQOL Speech. In addition to this, it ranks the very audible degradations produced by the GAN stego methods moderately stricter than DMOS, rather than the far less strict ratings of the other MOS-LQO algorithms. We would argue that this stricter rating is desirable when it comes to evaluating the perceptual transparency of audio steganography methods, as you would probably rather want a stricter rating than a more lenient one when it comes to technology used to secure secret information.

Stego method	MAE^{PESQ_F}	MAE^{PESQ_M}	MAE^{PESQ}
Steghide	0.3638	0.2990	0.3314
Hide4PGP	0.4857	0.2820	0.3838
GAN Low	1.6388	1.5643	1.6015
GAN High	1.7318	1.4750	1.6034
All Methods	1.0550	0.9051	0.9800

Table 4.6: The mean absolute errors (MAE) between DMOS and PESQ for female, male and all samples.

Stego method	$MAE^{ViSQOL-S_F}$	$MAE^{ViSQOL-S_M}$	$MAE^{ViSQOL-S}$
Steghide	0.2078	0.2018	0.2048
Hide4PGP	0.3440	0.1754	0.2596
GAN Low	1.2410	1.7448	1.4929
GAN High	1.3193	1.6506	1.4850
All Methods	0.7780	0.9431	0.8606

Table 4.7: The mean absolute errors (MAE) between DMOS and ViSQOL Speech for female, male and all samples.

Stego method	$MAE^{ViSQOL-A_F}$	$MAE^{ViSQOL-A_M}$	$MAE^{ViSQOL-A}$
Steghide	0.4599	0.3942	0.4270
Hide4PGP	0.5757	0.3711	0.4734
GAN Low	0.3852	0.4025	0.3939
GAN High	0.2812	0.5378	0.4095
All Methods	0.4255	0.4264	0.4259

Table 4.8: The mean absolute errors (MAE) between DMOS and ViSQOL Audio for female, male and all samples.

4.4 SNR Results

This section showcases the SNR results from our study. The SNR results measured for all audio samples degraded by the stego methods GAN Low and GAN High, along with the averages across all samples, all female samples and all male samples can be seen in Table 4.9, along with the DMOS scores from these same samples for comparison. The sample labels like "F1-S1" in Table 4.9 work like this: M and F refers to male and female speakers, the number behind this letter serves to differentiate between the two speakers per gender. The S stands for sample and serves to label the two samples spoken by each speaker. $AVG_{F/M/All}$ shows the average SNR scores across all samples, all female samples and all male samples. The SNR measurements for Steghide and Hide4PGP in a similarly formatted table can be found in Appendix F.

The reason why only the two GAN methods are shown in this section is because these are the results needed to dispute a claimed SNR threshold for human perception at 30 dB, which we found in [8] and [3]. We are disputing this threshold by comparing the SNR scores of these samples to their respective DMOS scores. By doing this we can check if the sample is audible to human perception and see if the SNR threshold at 30 dB holds up. It is important to remember that higher SNR scores are supposed to translate to better audio quality [3], i.e. harder to pick up by human perception.

The 30 dB SNR threshold is strongly disputed by our data, as all GAN method samples achieve DMOS scores below 3 and SNR scores above 34 dB. This already

strongly indicates that the threshold likely cannot be trusted, but when looking at some specific samples this indication gets even stronger. Sample F1-S2 achieves an SNR score of 41 dB, which is way above the claimed SNR threshold for human perception at 30 dB, along with a DMOS score of 2.6316 which, if we look at the DCR rating scale in Figure 3.4, translates to being somewhere between slightly annoying and annoying for our participants to listen to. Looking further at Table 4.9 we can also find other similar cases to this for the samples F1-S1, F2-S1 and F2-S2.

When comparing the average results from male and female samples in Table 4.9, we can also see that the human participants rates the male samples as having better quality on average than the female samples, while the SNR values suggest higher average quality for the female samples and lower for the male samples. This means that SNR and DMOS predicted opposite changes in audio quality between the male and female samples. Since DMOS is based on human perception [5], this could indicate that SNR might not align well with human perception, and raises a question about its suitability for evaluating the perceptual transparency of audio steganography methods altogether.

Sample	SNR^{GANLow}	$DMOS^{GANLow}$	$SNR^{GANHigh}$	$DMOS^{GANHigh}$
F1-S1	41.8081	2.8947	41.8448	2.6842
F1-S2	41.0197	2.6316	41.0461	2.5790
F2-S1	40.4620	2.4211	40.4922	2.2632
F2-S2	40.0506	2.3158	40.0529	2.3158
M1-S1	38.9550	2.9474	39.0194	3.0526
M1-S2	37.3718	2.7368	37.3833	3.0000
M2-S1	35.7744	2.5790	35.6228	2.8421
M2-S2	34.6223	2.3684	34.5545	2.1053
AVG_F	40.8351	2.5658	40.8590	2.4605
AVG_M	36.6809	2.6579	36.6450	2.7500
AVG_{All}	38.7580	2.6118	38.7520	2.6053

Table 4.9: Our measured SNR results for all GAN Low and GAN High samples, along with the averages for female samples, male samples and all samples.

4.5 Mathematical Error in TAN Based Method

A mathematical error is discovered when implementing the logistic tan map based audio steganography method described in [19]. An error where a "+1" to offset the iterations is omitted from one of the two functions that are part of a two dimensional logistic tan map used in the method is found. These functions can be seen to the right of Figure 3.3, and a similar two dimensional sine map function from [22] that led us to discover this error can be seen on the left, with the "+1"

omitted from the tan map function marked with red marker. More details about this error and exactly how it was found it can be seen in Section 3.5.4 of the method chapter.

Chapter 5

Discussion, Conclusions and Future Work

This chapter discusses the implications of our results, and uses these discussions to draw the conclusions used to answer our research questions. It starts by concluding which one of our three tested MOS-LQO algorithms we deem to be the most suited for evaluating the perceptual transparency of audio steganography methods, before moving on to dispute the 30 dB SNR threshold identified in previous work [3, 8], and questioning the suitability of SNR for measuring the perceptual transparency of audio steganography methods altogether. The chapter also mentions some potential limitations of our study and discusses how likely these are to have affected our results. The chapter also proposes some potential future work based on different observations we made while working on this thesis.

5.1 Most Suited MOS-LQO Algorithm

In this section we will discuss which one of our tested MOS-LQO Algorithms; PESQ, ViSQOL Speech and ViSQOL Audio that appears to be the most suited for evaluating the perceptual transparency metric in audio steganography, in doing this we will answer our first research question: "How do the different MOS-LQO algorithms; PESQ, ViSQOL Speech, and ViSQOL Audio compare to a subjective DMOS test, when it comes to evaluating the perceptual transparency of our chosen audio steganography methods, and which one appears to be the best suited?", as defined in Section 1.2.

Judging from our findings listed in Section 4.3, comparing our MOS-LQO and Subjective DMOS results, it is quite clear to us that ViSQOL Audio appears to be by far the best suited MOS-LQO algorithm for evaluating the perceptual transparency of our chosen audio steganography methods. We say this, firstly because ViSQOL Audio correlates the closest to DMOS in terms of Pearson's correlation, with PESQ coming in at a close second, and ViSQOL Speech falling quite far behind the others. And secondly because, ViSQOL Audio deviates way less from the

others in terms of our Mean Absolute Error (MAE) measurements, deviating about three times less than the others. Thirdly, we also saw in our visual analysis of the scatter plots and raw MAE data that ViSQOL Audio appears to have a tendency to evaluate the perceptual transparency of the tested samples moderately more strictly than DMOS, for the samples with the most audible quality degradations according to DMOS. Meanwhile, PESQ and ViSQOL Speech had a tendency to evaluate considerably less strict than DMOS. We would argue that this moderate strictness of ViSQOL Audio is a clear advantage over considerable lenience observed in PESQ and ViSQOL Speech for large degradations, when it comes to picking a suitable MOS-LQO algorithm to evaluate perceptual transparency in a security focused field like audio steganography.

However, in defense of the other MOS-LQO algorithms, one could argue that their scores would still have been bad enough that a security conscious actor using them to evaluate our tested GAN methods in particular, which where the methods with the largest degradations in audio quality, would have seen that the method is indeed perceptible and kept looking for alternatives. However, the DMOS test and manual observations from the author suggests that the GAN methods in question were extremely easy to hear. Our two other stego methods; Steghide and Hide4PGP, achieved DMOS scores close to the null-pairs, suggesting that they are either impossible or extremely difficult to perceive by a human observer. It would therefore have been interesting to see future work where our tested MOS-LQO algorithms are compared in evaluating some steganography methods where the perceptibility to humans is more borderline, where some of the participants will hear a clear degradation and some wont, to see what algorithms are able to differentiate these borderline cases. For example testing a steganography method achieving a DMOS score slightly above 4.

In addition to this, ViSQOL Audio also has another clear benefit, being that it works on all types of audio samples. In contrast to this, PESQ and ViSQOL Speech can only evaluate the perceptual transparency of speech samples. This is another distinct advantage of the ViSQOL Audio algorithm, as it makes it more flexible, and potentially allows it to evaluate more audio steganography methods than the other tested MOS-LQO algorithms, such as methods made specifically for embedding information in music. This may prove to be especially beneficial for testing audio watermarking [112] methods made for the music industry.

Yet another benefit of the ViSQOL Audio algorithm, that it also shares with ViSQOL Speech, is that it is completely free to use and open source. In Section 2.6.1 we mention the licensing issues of PESQ, and how we interpret its license terms to not allow for unlicensed evaluation of audio steganography methods, even academically, unless the main purpose of the study is to evaluate the algorithm itself and not the audio steganography methods. While ViSQOL Speech shares this benefit, it performs the worst our of our tested MOS-LQO algorithms overall, getting similar MAE results to PESQ and quite a bit worse correlation results than the similar results of ViSQOL Speech and PESQ.

Judging from our results and this discussion we deem that ViSQOL Audio is

very likely the most suited, out of our three tested MOS-LQO algorithms, for evaluating the perceptual transparency of audio steganography methods. It performs by far the closest to our subjective DMOS test, with Subjective MOS tests like this often being considered the ground truth for MOS testing [12], as well as appearing to have both better license terms than PESQ and increased flexibility over both PESQ and ViSQOL Speech, by allowing the testing of non-speech samples. While more work could be beneficial to see if it also performs this well for non-speech samples, and for other audio steganography methods than tested in this thesis, we still think that the audio steganography field could benefit from replacing PESQ with ViSQOL Audio as the "go-to" MOS-LQO algorithm for perceptual transparency testing of audio steganography methods.

5.2 Previously Assumed SNR Threshold Disputed

In this section we will answer our second research question; "How do the SNR scores compare to the subjective DMOS scores from our experiments, do the results support or oppose a threshold of 30 dB for human perception?", as defined in Section 1.2.

Judging from our signal to noise ratio (SNR) results in Section 4.4, we can clearly see that the SNR threshold for human perception identified in the literature [3, 8], claiming that audio with SNR scores above 30 dB is inaudible to humans does not hold up. This is clear as all tested audio samples achieve SNR values higher than 30 dB, while simultaneously achieving DMOS scores that suggest that the audio samples are between slightly annoying and annoying to listen to for a human observer. We also discuss some even more extreme cases for single samples in Section 4.4 of the Results chapter. Future work exploring whether such an SNR threshold for human perception could make sense, and if it does trying to identify one, could perhaps also be interesting.

5.3 Questioning SNR's Suitability for Perceptual Transparency Testing

Another interesting finding from our SNR results in Section 4.4 is that the DMOS scores from the male and female samples from the GAN High method indicate that the male samples have moderately higher audio quality than the female samples, while the SNR values indicate the opposite. This inverse correlation makes us wonder whether SNR is suited for perceptual transparency testing at all. The reason why we scrutinize SNR rather than DMOS in this case is that SNR is not based on human perception at all [17], while DMOS is a subjective test based entirely on ratings from human participants [5].

We can also think of other cases where using SNR as a metric for perceptual transparency could potentially cause problems. While surveying the literature, we

came across a subset category of audio steganography methods exploiting weaknesses of the human auditory system (HAS), often referred to as tone insertion methods [2], such as [113] which inserts high pitch tones directly into an audio file to embed a secret message. We wonder how SNR would work in evaluating a method like this, and if the high pitch tones will be picked up as noise, potentially resulting in a lower SNR score and giving the impression of poor perceptual transparency, even though this high pitch tone might not be perceptible to human listeners.

Because of the things discussed in this section, we would like to see future work exploring the suitability of SNR for perceptual transparency testing withing the field of audio steganography as a whole. We would also like to see future work exploring how suited SNR is to evaluate the perceptual transparency of audio steganography methods utilizing tone insertion in particular, preferably including at least one method utilizing high pitch tone insertion.

5.4 Potential Limitations of Our Study

5.4.1 General Limitations of Subjective Testing

Every subjective test like the DMOS test done in this study will likely have variations due to the large amount of factors that can affect them. A good example of this is illustrated in Figure 3.1 showing a plethora of things that can affect subjective MOS testing for evaluating speech quality. This paper [55] on MOS limitations also shows examples of MOS testing for video quality from different labs causing variations in the results. It is possible that that the fact that we moved around to different locations during our DMOS tests have caused variations in the results. However, we did have measures in place, such as using the same equipment, noise-canceling headphones, and as quiet of an environment as possible at each location to try to limit these variations. Even so, this inherent weakness of subjective studies could call for future work trying to reproduce our results, in order to further ensure that ViSQOL Audio is the best of our three tested MOS-LQO at assessing the perceptual transparency of audio steganography methods.

However, our results are also so strongly in favor of ViSQOL Audio that we find it unlikely that some potential randomness in the DMOS scores, caused by inherent possible variations of a subjective study like this could account for the entire benefit towards ViSQOL Audio. We also saw several other studies using somewhere around 20 participants for MOS testing [4, 58], and we think that this is likely enough to get a valid result. We would there still absolutely suggest that the audio steganography field change from PESQ to ViSQOL Audio as the "go-to" MOS-LQO algorithms for perceptual transparency testing within audio steganography, despite this potential limitation. Further work replicating our work with a DMOS test including more participants could probably also be beneficial to further confirm the validity our results.

5.4.2 Potential Limitations of Our Chosen Audio Samples

While we tried picking high quality reference audio files with as little disturbances as possible for our testing, some of our audio samples did still have some small barely noticeable disturbances. A considerable amount of time was spent trying to find a suitable dataset, and our audio samples were simply the best we were able to find within the time-frame of this thesis, that also fulfilled the requirements of our ITU-T P.800 [5] DCR DMOS test.

We did however try to compensate for these small disturbances in some of our audio samples, both by using DCR over ACR, so that our participants has a chance to hear that the disturbances is present in both samples, hopefully causing them to leave this out of their audio degradation rating, and by choosing to present each pair of high quality reference and degraded audio samples twice, so that the participants has several chances to hear if a degradation is present in both samples or just the degraded one. We think that these choices likely compensated for this limitation to a high degree, but we still think it is possible that the results could have been affected to some degree by these small degradations in the high quality reference audio samples. We do not find it unlikely that some of our participants may have rated certain samples slightly differently as a result of not noticing that a degradation was present in both samples. However, we do not think it is likely to have affected our final DMOS scores to a considerable degree.

We did however, notice a pattern in our raw DMOS data (that can be found on our GitHub [67] "Supplementary Materials" directory.) where one sample (Sample M1-S2) achieved a perfect DMOS score (5) across the null-pairs, Steghide and Hide4PGP. Meaning that all the non-excluded participants gave this sample a perfect score across all of these three degradations. When listening to this file over and over, it appears to be nearly perfect quality with zero audible degradations to the author's ears. This pattern of having a perfect score across all of these three degradations is interesting, and could further suggest that the participants of our test sometimes rated degradations present in both the high quality reference and degraded audio samples more harshly than sample pairs that did not have these degradations in their reference files.

It is also possible that the way the speaker of this sample sounded and/or pronounced his words was just somehow very pleasant to the listeners, and that this made them ignore degradations that may actually have been there for Steghide and Hide4PGP. The fact that this sample scores slightly higher than average for our GAN Low samples and a decent amount higher for GAN High could support this theory to some degree.

We also think that having the null-pair grounding will account for this potential limitation to a high degree, by showing us how the participants rate all of our high quality samples against themselves. As explained in the ITU-T P.800 [5], the null-pair test serves as a sort of grounding of the scores, showing us what scores we can expect for a perfect sample. While this likely does not address the entire limitation, if it is even present, we still think addresses it to a high degree.

5.4.3 Somewhat Extreme Steganography Method DMOS Scores

As briefly mentioned in Section 5.1, another potential limitation of our study could be the somewhat extreme results of our tested audio steganography methods. Our Steghide and Hide4PGP methods achieve very high DMOS scores, that come close to our null-pair results, suggesting that they are likely imperceptible or extremely hard to hear for most people. Meanwhile, our GAN methods both end up with average DMOS scores that fall between "annoying" and "slightly annoying" on the DCR scale, that can be seen in Figure 3.4.

While it can be difficult to know how audio steganography methods perform beforehand, and Reyer's thesis [3] conveys having difficulties finding recently proposed audio steganography methods that can easily be implemented from their papers, future work testing how different MOS-LQO algorithms perform in methods that produce DMOS scores that are more on the borderline of human perception could be interesting. We think that this is likely an important area to get right for MOS-LQO algorithm, to avoid methods being labeled as imperceptible while they may actually still be perceptible to a decently large part of the population. We think that a method achieving a DMOS score of slightly above 4, or slightly above "audible but not annoying" on the DCR scale (Figure 3.4) would likely fit these criteria.

Bibliography

[1] United Nations, *THE 17 GOALS | Sustainable Development*. [Online]. Available: https://sdgs.un.org/goals (visited on 09/06/2025).

[2] A. A. AlSabhany, A. H. Ali, F. Ridzuan, A. H. Azni and M. R. Mokhtar, 'Digital audio steganography: Systematic review, classification, and analysis of the current state of the art,' *Computer Science Review*, vol. 38, p. 100 316, Nov. 2020, ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2020.100316. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013720304160 (visited on 21/01/2025).

[3] P. M. Reyers, *A comparative analysis of audio steganography methods and tools*, Publisher: University of Twente, Jul. 2023. [Online]. Available: https://essay.utwente.nl/95988/ (visited on 21/01/2025).

[4] S. S. Bharti, M. Gupta and S. Agarwal, 'A novel approach for audio steganography by processing of amplitudes and signs of secret audio separately,' en, *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 23 179–23 201, Aug. 2019, ISSN: 1573-7721. DOI: 10.1007/s11042-019-7630-4. [Online]. Available: https://doi.org/10.1007/s11042-019-7630-4 (visited on 23/04/2025).

[5] ITU Telecommunication Standardization Sector, *P.800 : Methods for subjective determination of transmission quality*, Aug. 1996. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I (visited on 20/01/2025).

[6] ITU Telecommunication Standardization Sector, *P.800.1: Mean opinion score (MOS) terminology*, Jul. 2016. [Online]. Available: https://www.itu.int/rec/T-REC-P.800.1-201607-I/en (visited on 22/01/2025).

[7] OPTICOM GmbH, *Voice Quality Testing | PESQ*. [Online]. Available: https://www.opticom.de/technology/pesq.php (visited on 01/05/2025).

[8] S. Hemalatha and Ramathmika, 'A Robust MP3 Audio Steganography with Improved Capacity,' in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Oct. 2020, pp. 640–645. DOI: 10.1109/ICCCA49541.2020.9250894. [Online]. Available: https://ieeexplore.ieee.org/document/9250894 (visited on 01/05/2025).

[9] H. H. Stormyhr, 'IMT4205 - Research Project Planning report "Modern-day steganography" [Unpublished master's thesis pre-project report from the course IMT4205 at NTNU],' *NTNU*, Dec. 2024.

[10] OPTICOM GmbH, *Perceptual voice perceptual audio quality test*. [Online]. Available: https://www.opticom.de/technology/peaq.php (visited on 06/05/2025).

[11] ITU Telecommunication Standardization Sector, *P.863 : Perceptual objective listening quality prediction*, Mar. 2018. [Online]. Available: https://www.itu.int/rec/t-rec-p.863 (visited on 22/01/2025).

[12] A. Hines, J. Skoglund, A. C. Kokaram and N. Harte, 'ViSQOL: An objective speech quality model,' en, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, May 2015, ISSN: 1687-4722. DOI: 10.1186/s13636-015-0054-9. [Online]. Available: https://doi.org/10.1186/s13636-015-0054-9 (visited on 22/01/2025).

[13] Sevana Oü, *AQuA - Audio Quality Analyzer alternative for POLQA and ViSQOL*, en-US. [Online]. Available: https://sevana.biz/products-aqua-polqa-visqol/ (visited on 22/01/2025).

[14] ludlows, *PESQ Intellectual Property Rights Notice*, en. [Online]. Available: https://github.com/serser/python-pesq/blob/master/pypesq/pesq.h (visited on 06/05/2025).

[15] A. Rix, J. Beerends, M. Hollier and A. Hekstra, 'Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,' in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023. [Online]. Available: https://ieeexplore.ieee.org/document/941023 (visited on 06/05/2025).

[16] C.-S. Lin and K.-S. Bae, 'Correlation Analysis of PESQ and MOS Evaluation for HMM-based Synthetic Korean Speech,' kor, *Phonetics and Speech Sciences*, vol. 2, no. 1, pp. 71–75, 2010, ISSN: 2005-8063. [Online]. Available: https://koreascience.kr/article/JAKO201019455946036.page (visited on 22/04/2025).

[17] P. Lechner, *SNR/SNR.ipynb at main · hrtlacek/SNR*, en, Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4724137 (visited on 06/05/2025).

[18] ITU Telecommunication Standardization Sector, *P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Feb. 2001. [Online]. Available: https://www.itu.int/rec/t-rec-p.862 (visited on 22/01/2025).

[19] M. T. Elkandoz and W. Alexan, 'Logistic Tan Map Based Audio Steganography,' in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, Nov. 2019, pp. 1–5. DOI: 10.1109/ICECTA48151.2019.8959683. [Online]. Available: https://ieeexplore.ieee.org/document/8959683 (visited on 07/05/2025).

[20] Microsoft Corporation, *Microsoft Copilot: Your AI companion*, en-us. [Online]. Available: https://copilot.microsoft.com (visited on 31/01/2025).

[21] Z. Hua, Y. Zhou, C.-M. Pun and C. L. P. Chen, 'Image encryption using 2D Logistic-Sine chaotic map,' in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2014, pp. 3229–3234. DOI: 10.1109/SMC.2014.6974425. [Online]. Available: https://ieeexplore.ieee.org/document/6974425 (visited on 15/05/2025).

[22] M. T. Elkandoz, W. Alexan and H. H. Hussein, 'Logistic Sine Map Based Image Encryption,' in *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Sep. 2019, pp. 290–295. DOI: 10.23919/SPA.2019.8936718. [Online]. Available: https://ieeexplore.ieee.org/document/8936718 (visited on 15/05/2025).

[23] NTNU, *Sustainability in Computing Education - NTNU*. [Online]. Available: https://www.ntnu.edu/web/excited/sustainability-in-computing-education (visited on 03/06/2025).

[24] United Nations, *Goal 9 | Department of Economic and Social Affairs*. [Online]. Available: https://sdgs.un.org/goals/goal9#targets_and_indicators (visited on 09/06/2025).

[25] United Nations, *Goal 16 | Department of Economic and Social Affairs*. [Online]. Available: https://sdgs.un.org/goals/goal16#targets_and_indicators (visited on 09/06/2025).

[26] Wikipedia, *Steganography*, en, Apr. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Steganography&oldid=1287922694 (visited on 07/05/2025).

[27] Wikipedia, *Jeremiah Denton*, en, May 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Jeremiah_Denton&oldid=1288832539 (visited on 07/05/2025).

[28] Wikipedia, *List of steganography techniques*, en, Mar. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_steganography_techniques&oldid=1282887792 (visited on 07/05/2025).

[29] L. Craig, *What is AI watermarking and how does it work?* en, Oct. 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/AI-watermarking (visited on 15/05/2025).

[30] S. Hetzl, *Steghide*, Oct. 2003. [Online]. Available: https://steghide.sourceforge.net/index.php (visited on 08/05/2025).

[31] S. De Vuono, *Steghide man page*, en, Oct. 2003. [Online]. Available: https://github.com/StegHigh/steghide/blob/master/doc/steghide.1 (visited on 08/05/2025).

[32] H. Repp, *Hide4PGP Homepage*, Feb. 2000. [Online]. Available: https://web.archive.org/web/20240406191720/http://www.heinz-repp.onlinehome.de/Hide4PGP.htm (visited on 08/05/2025).

[33] Wikipedia, *Bit numbering*, en, Apr. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bit_numbering&oldid=1283658568#Least_significant_bit (visited on 08/05/2025).

[34] P. Mahajan, *Steganography: A Data Hiding Technique*, en, Nov. 2014. [Online]. Available: https://www.researchgate.net/publication/344412253_Steganography_A_Data_Hiding_Technique (visited on 08/05/2025).

[35] G. Tzanetakis, G. Essl and P. Cook, *Automatic Musical Genre Classification Of Audio Signals*, 2001. [Online]. Available: https://huggingface.co/datasets/marsyas/gtzan/blob/main/README.md (visited on 08/05/2025).

[36] Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor and Fiscus, Jonathan G., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993. DOI: 10.35111/17GK-BN40. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1 (visited on 21/01/2025).

[37] D. Ye, S. Jiang and J. Huang, *Heard More Than Heard: An Audio Steganography Method Based on GAN*, Jul. 2019. DOI: 10.48550/arXiv.1907.04986. [Online]. Available: http://arxiv.org/abs/1907.04986 (visited on 08/05/2025).

[38] K. A. Zhang, A. Cuesta-Infante, L. Xu and K. Veeramachaneni, *SteganoGAN: High Capacity Image Steganography with GANs*, Jan. 2019. DOI: 10.48550/arXiv.1901.03892. [Online]. Available: http://arxiv.org/abs/1901.03892 (visited on 14/05/2025).

[39] A. Garg and S. Rosetti, *Implementation of StegoGAN: High Capacity Image Steganography for Image and Audio Input with GANs*, Jul. 2020. [Online]. Available: https://github.com/garg-akash/Steganography_GANs (visited on 14/05/2025).

[40] Wikipedia, *Generative adversarial network*, en, Apr. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Generative_adversarial_network&oldid=1284548497 (visited on 14/05/2025).

[41] Wikipedia, *Advanced Encryption Standard*, en, May 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Advanced_Encryption_Standard&oldid=1290218924 (visited on 14/05/2025).

[42] The POLQA Coalition, *POLQA - The Next-Generation Mobile Voice Quality Testing Standard*. [Online]. Available: http://www.polqa.info/ (visited on 16/05/2025).

[43] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, 'Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model,' en, *Journal of the audio engineering society*, vol. 50, no. 10, pp. 765–778, Oct. 2002, Alternative URL: http://www.mp3-tech.org/programmer/docs/2001-P03b.pdf. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=11062 (visited on 28/01/2025).

[44] ITU Telecommunication Standardization Sector, *P.862.1 : Mapping function for transforming P.862 raw result scores to MOS-LQO*, Nov. 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.1-200311-W/en (visited on 16/05/2025).

[45] ITU Telecommunication Standardization Sector, *P.862.2 : Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, Nov. 2007. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.2-200711-W/en (visited on 16/05/2025).

[46] ludlows, *GitHub - ludlows/PESQ at v0.0.4*, en, May 2019. [Online]. Available: https://github.com/ludlows/PESQ (visited on 16/05/2025).

[47] S. Terpinas and G. Zachos, *Google/visqol*, Dec. 2019. [Online]. Available: https://github.com/google/visqol (visited on 16/05/2025).

[48] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman and A. Hines, *ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric*, Apr. 2020. DOI: 10.48550/arXiv.2004.09584. [Online]. Available: http://arxiv.org/abs/2004.09584 (visited on 16/05/2025).

[49] OPTICOM GmbH, *OPTICOM voice quality PESQ PEXQ speech quality perceptual measurement*. [Online]. Available: https://www.opticom.de/ (visited on 06/05/2025).

[50] The POLQA Coalition, *POLQA - The Next-Generation Mobile Voice Quality Testing Standard*. [Online]. Available: http://www.polqa.info/products.html (visited on 16/05/2025).

[51] Sevana Oü, *Call quality testing & monitoring. AQuA PVQA POLQA ViSQOL*, en-US. [Online]. Available: https://sevana.biz/ (visited on 07/06/2025).

[52] OPTICOM GmbH, *PEAQ BS.1387 BS.1116 Perceptual Evaluation of Audio Quality*. [Online]. Available: https://www.opticom.de/licensing/peaq.php (visited on 16/05/2025).

[53] ITU Telecommunication Standardization Sector, *P.863.1 : Application guide for Recommendation ITU-T P.863*, Jun. 2019. [Online]. Available: https://www.itu.int/rec/T-REC-P.863.1-201906-I/en (visited on 07/06/2025).

[54] DeepL SE, *Seamless PDF translation with unmatched accuracy*, en. [Online]. Available: https://www.deepl.com/en/features/document-translation/pdf (visited on 22/04/2025).

[55] R. C. Streijl, S. Winkler and D. S. Hands, 'Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives,' en, *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, Mar. 2016, ISSN: 1432-1882. DOI: 10.1007/s00530-014-0446-1. [Online]. Available: https://doi.org/10.1007/s00530-014-0446-1 (visited on 29/01/2025).

[56] ITU Telecommunication Standardization Sector, *P.911 : Subjective audiovisual quality assessment methods for multimedia applications*, Dec. 1998. [Online]. Available: https://www.itu.int/rec/T-REC-P.911-199812-W/en (visited on 23/04/2025).

[57] ITU Telecommunication Standardization Sector, *BT.500 : Methodologies for the subjective assessment of the quality of television images*, May 2023. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500 (visited on 23/04/2025).

[58] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely and J. Gustafson, 'Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation,' in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 41–47. DOI: 10.21437/SSW.2023-7. [Online]. Available: https://www.isca-archive.org/ssw_2023/kirkland23_ssw.html# (visited on 23/04/2025).

[59] Sikt, *Avtaler om personverntjenester for forskning | Sikt*, nb. [Online]. Available: https://sikt.no/tjenester/personverntjenester-forskning/avtaler-om-personverntjenester-forskning (visited on 24/04/2025).

[60] Sikt, *Meldeskjema for behandling av personopplysninger*. [Online]. Available: https://meldeskjema.sikt.no/test (visited on 24/04/2025).

[61] Sikt, *Information for participants in research projects | Sikt*, en. [Online]. Available: https://sikt.no/en/tjenester/personverntjenester-forskning/fylle-ut-meldeskjema-personopplysninger/information-participants-research-projects (visited on 24/04/2025).

[62] A. Wrench, 'MOCHA-TIMIT,' *Queen Margaret University College*, Nov. 1999. [Online]. Available: https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (visited on 21/01/2025).

[63] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, 'PTDB-TUG: Pitch Tracking Database from Graz University of Technology — SPSC @ TU Graz,' Aug. 2011. [Online]. Available: https://www.spsc.tugraz.at/databases-and-tools/ptdb-tug-pitch-tracking-database-from-graz-university-of-technology.html (visited on 21/01/2025).

[64] J. Kominek and A. W. Black, *CMU ARCTIC databases for speech synthesis*, 2003. [Online]. Available: http://festvox.org/cmu_arctic/ (visited on 22/04/2025).

[65] S. Jacob, *Mean Opinion Scores (MOS) | Voice Quality Online Seminar*, HEAD acoustics International - YouTube, Oct. 2019. [Online]. Available: https://www.youtube.com/watch?v=pj3DBj2FSlw (visited on 23/01/2025).

[66] Wikipedia, *Intelligibility (communication)*, en, Page Version ID: 1213932495, Mar. 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Intelligibility_(communication)&oldid=1213932495 (visited on 23/01/2025).

[67] H. H. Stormyhr, *HHS Master's Thesis GitHub Repo*, Jun. 2025. [Online]. Available: https://github.com/HenrikHStormyhr/HHS-Masters-Thesis (visited on 09/06/2025).

[68] OffSec Services Limited, *Get Kali*, English. [Online]. Available: https://www.kali.org/get-kali/ (visited on 20/05/2025).

[69] Oracle, *Downloads – Oracle VirtualBox*. [Online]. Available: https://www.virtualbox.org/wiki/Downloads (visited on 20/05/2025).

[70] gbmb.org, *KB to Bytes Conversion Kilobytes to Bytes Calculator*. [Online]. Available: https://www.gbmb.org/kb-to-bytes (visited on 20/05/2025).

[71] lipsum.com, *Lorem Ipsum - All the facts - Lipsum generator*. [Online]. Available: https://www.lipsum.com/ (visited on 20/05/2025).

[72] M. Wojciakowski et al., *What is Windows Subsystem for Linux*, en-us. [Online]. Available: https://learn.microsoft.com/en-us/windows/wsl/about (visited on 20/05/2025).

[73] R. B. Dannenberg, D. Mazzoni and Muse_Group, *Audacity ® | Free Audio editor, recorder, music making and more!* [Online]. Available: https://www.audacityteam.org/ (visited on 21/01/2025).

[74] OpenAI, *Introducing ChatGPT*, en-US, Nov. 2022. [Online]. Available: https://openai.com/index/chatgpt/ (visited on 23/05/2025).

[75] D. Tan, Y. Lu, X. Yan and X. Wang, 'A Simple Review of Audio Steganography,' in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Mar. 2019, pp. 1409–1413. DOI: 10.1109/ITNEC.2019.8729476. [Online]. Available: https://ieeexplore.ieee.org/document/8729476/?arnumber=8729476&tag=1 (visited on 02/10/2024).

[76] Wikipedia, *Latin square*, en, Feb. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Latin_square&oldid=1273465655 (visited on 28/04/2025).

[77] Fiveable, *Replication, randomization, and local control | Experimental Design Class Notes*, en. [Online]. Available: https://library.fiveable.me/experimental-design/unit-2/replication-randomization-local-control/study-guide/vIxW7WpYfRK9vDp0 (visited on 28/04/2025).

[78] D. Masson, *Balanced Latin Square Online Generator*, en. [Online]. Available: https://damienmasson.com/tools/latin_square/ (visited on 28/04/2025).

[79] Power Norge AS, *Eletra CA101 Chat 3,5mm headset*, no. [Online]. Available: https://www.power.no/data-og-tilbehoer/pc-lyd/pc-hodetelefoner/eletra-ca101-chat-35mm-headset/p-2362806/ (visited on 01/05/2025).

[80] OPTICOM GmbH, *What is "POLQA"?* [Online]. Available: https://www.opticom.de/technology/polqa.php (visited on 06/05/2025).

[81] Microsoft Corporation, *Visual Studio Code - Code Editing. Redefined*, en. [Online]. Available: https://code.visualstudio.com/ (visited on 23/05/2025).

[82] Microsoft Corporation, *Microsoft C++ Build Tools*, en-US. [Online]. Available: https://visualstudio.microsoft.com/visual-cpp-build-tools/ (visited on 23/05/2025).

[83] Charles R Harris, Matthew Brett, Matti Picus, Ralf Gommers and Travis Oliphant, *Numpy: Fundamental package for array computing in Python*.

[84] Google LLC, *Installing Bazel on Windows*, en, Apr. 2025. [Online]. Available: https://bazel.build/install/windows (visited on 24/05/2025).

[85] mahyoseung and rsanchezpizani, *Armadillo is not found (error 404) · Issue #126 · google/visqol*, en, Dec 12, 2024. [Online]. Available: https://github.com/google/visqol/issues/126 (visited on 24/05/2025).

[86] C. Sanderson and R. Curtin, *Armadillo Files*, Oct. 2024. [Online]. Available: https://sourceforge.net/projects/arma/files/ (visited on 24/05/2025).

[87] P. M. Reyers, *Steganography-analysis/compute_transparency.py at main*, en, Jun. 2023. [Online]. Available: https://github.com/MatthijsReyers/steganography-analysis/blob/main/compute_transparency.py (visited on 23/05/2025).

[88] GeeksforGeeks, *Confidence Interval*, en-US, Feb. 2025. [Online]. Available: https://www.geeksforgeeks.org/confidence-interval/ (visited on 26/05/2025).

[89] GeeksforGeeks, *P-Value: Comprehensive Guide to Understand, Apply, and Interpret*, en-US, Jan. 2024. [Online]. Available: https://www.geeksforgeeks.org/p-value/ (visited on 08/06/2025).

[90] A. Srivastava, *Detecting Anomalies with Z-Scores: A Practical Approach*, en, Oct. 2023. [Online]. Available: https://medium.com/@akashsri306/detecting-anomalies-with-z-scores-a-practical-approach-2f9a0f27458d (visited on 24/05/2025).

[91] S. Bradburn, *How To Calculate Z Scores In Excel*, Feb. 2020. [Online]. Available: https://www.youtube.com/watch?v=oeljG6kU5S4 (visited on 24/05/2025).

[92] S. Bradburn, *How To Perform T-Tests In Microsoft Excel*, Jan. 2019. [Online]. Available: https://www.youtube.com/watch?v=q0ckcKsSPXU (visited on 24/05/2025).

[93] Wikipedia, *Paired difference test*, en, May 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Paired_difference_test&oldid=1290349570 (visited on 26/05/2025).

[94] JMP Statistical Discovery LLC, *Paired t-Test*, en. [Online]. Available: https://www.jmp.com/en/statistics-knowledge-portal/t-test/paired-t-test (visited on 25/05/2025).

[95] G. G. Løvås, *Statistikk for universitetet og høgskoler*, no, 4th ed. Oslo, Norway: Universitetsforlaget, 2021, ISBN: 978-82-15-03104-0.

[96] G. Malato, *An Introduction to the Shapiro-Wilk Test for Normality | Built In*, Jan. 2025. [Online]. Available: https://builtin.com/data-science/shapiro-wilk-test (visited on 25/05/2025).

[97] Statistics Kingdom, *Shapiro-Wilk test calculator: Normality calculator, Q-Q plot*. [Online]. Available: https://www.statskingdom.com/shapiro-wilk-test-calculator.html (visited on 25/05/2025).

[98] University of Southampton, *T test | Practical Applications of Statistics in the Social Sciences | University of Southampton*. [Online]. Available: https://www.southampton.ac.uk/passs/gcse_scores/bivariate_analysis/t_test.page (visited on 25/05/2025).

[99] Social Science Statistics, *Wilcoxon Signed-Rank Test Calculator*. [Online]. Available: https://www.socscistatistics.com/tests/signedranks/ (visited on 25/05/2025).

[100] IntellectusConsulting, *Understanding the Wilcoxon Signed Rank Test*, en-US. [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/how-to-conduct-the-wilcox-sign-test/ (visited on 25/05/2025).

[101] Wikipedia, *Wilcoxon signed-rank test*, en, May 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wilcoxon_signed-rank_test&oldid=1291114696#Zeros_and_ties (visited on 25/05/2025).

[102] Social Science Statistics, *Sign Test Calculator*. [Online]. Available: https://www.socscistatistics.com/tests/signtest/default.aspx (visited on 26/05/2025).

[103] Wikipedia, *Pearson correlation coefficient*, en, May 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1290799627 (visited on 25/05/2025).

[104] Wikipedia, *Mean absolute error*, en, Feb. 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Mean_absolute_error&oldid=1276071917 (visited on 25/05/2025).

[105] The R Foundation, *R: The R Project for Statistical Computing*. [Online]. Available: https://www.r-project.org/ (visited on 25/05/2025).

[106] Posit Software, *Posit*, en. [Online]. Available: https://www.posit.co/ (visited on 25/05/2025).

[107] S. Bradburn, *How To Perform A Pearson Correlation Test In R*, Mar. 2020. [Online]. Available: https://www.youtube.com/watch?v=2J_ZlxLeuQU (visited on 30/05/2025).

[108] J. Jingze, *How to Use Excel to Calculate MAE and MSE based on Naive Forecasting Results*, May 2024. [Online]. Available: https://www.youtube.com/watch?v=lutysCSzKG0 (visited on 25/05/2025).

[109] The Data Visualisation Catalouge, *Box and Whisker Plot*. [Online]. Available: https://datavizcatalogue.com/methods/box_plot.html (visited on 08/06/2025).

[110] Statistics Kingdom, *Advanced box and whisker plot maker*. [Online]. Available: https://www.statskingdom.com/advanced-boxplot-maker.html (visited on 29/05/2025).

[111] Photopea, *Photopea | Online Photo Editor*. [Online]. Available: https://www.photopea.com/ (visited on 29/05/2025).

[112] G. Hua, J. Huang, Y. Q. Shi, J. Goh and V. L. L. Thing, 'Twenty years of digital audio watermarking—a comprehensive review,' *Signal Processing*, vol. 128, pp. 222–242, Nov. 2016, ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2016.04.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168416300263 (visited on 08/06/2025).

[113] A. Habes, 'HIDING INFORMATION IN WAV-FILE - Implementation, Analysis and Evaluation,' en, vol. 2, SCITEPRESS, Apr. 2006, pp. 274–281, ISBN: 978-972-8865-46-7. DOI: 10.5220/0001257802740281. [Online]. Available: https://www.scitepress.org/PublishedPapers/2006/12578 (visited on 02/06/2025).

Appendix A

Master Agreement

The master agreement for the project is attached below. This outlines the original project description. While this is largely unchanged the MOS-LQS term is somewhat inaccurate as we decided to move from ACR (typically labeled MOS-LQS) to DCR (typically labeled DMOS) during the project. Using the term DMOS instead of MOS-LQS would therefore be more accurate with this change in mind.

The master agreement is included in both Norwegian and English.

Masteravtale/hovedoppgaveavtale

Sist oppdatert 11. november 2020

Fakultet	**Fakultet for informasjonsteknologi og elektroteknikk**
Institutt	**Institutt for informasjonssikkerhet og kommunikasjonsteknologi**
Studieprogram	**MIS**
Emnekode	**MIS4900**

Studenten

Etternavn, fornavn	**Stormyhr, Henrik Hansen**
Fødselsdato	**30.09.1999**
E-postadresse ved NTNU	**henrhst@stud.ntnu.no**

Tilknyttede ressurser

Veileder	**Tjerand Silde**
Eventuelle medveiledere	**Emil August Hovd Olaisen, Bor de Kock**
Eventuelle medstudenter	

Oppgaven

Oppstartsdato	**06.01.2025**
Leveringsfrist	**10.06.2025**
Oppgavens arbeidstittel	**A comparative analysis of MOS-LQO algorithms for perceptual transparency testing in audio steganography**
Problembeskrivelse	**Perceptual transparency testing is highly relevant for evaluating audio steganography methods. The perceptual transparency of such methods says something about how easy it is for a human observer to hear that a secret message has been embedded into an audio file. One way to test this is by using Objective Mean Opinion Score (MOS-LQO) algorithms. These algorithms try to emulate Subjective Mean Opinion Score (MOS-LQS) tests where human participants rate the perceived quality of some altered/degraded audio files. However, there exists a plethora of different MOS-LQO algorithms and no current work is exploring which one is best suited for evaluating the perceptual transparency of different audio steganography methods. This study aims to change this by investigating how some different MOS-LQO algorithms compare to Subjective MOS (MOS-LQS) tests when evaluating the perceptual transparency of some chosen audio steganography tools and methods.**

Risikovurdering og datahåndtering

Skal det gjennomføres risikovurdering? **Dersom «ja», har det blitt gjennomført?**	**Nei** **Nei**
Skal det søkes om godkjenninger? (REK*, Sikt)**	**Ja**
Skal det skrives en konfidensialitetsavtale i forbindelse med oppgaven? **Hvis «ja», har det blitt gjort?**	**Nei** **Nei**

* Regionale komiteer for medisinsk og helsefaglig forskningsetikk (https://rekportalen.no)

** Sikts meldeskjema for personopplysninger i forskning (https://sikt.no/tjenester/personverntjenester-forskning/fylle-ut-meldeskjema-personopplysninger)

Eventuelle emner som skal inngå i mastergraden

IMT4110 Scientific Methodology and Communication IMT4115 Introduksjon til informasjonssikkerhetsledelse IMT4114 Introduction to Digital Forensics IMT4113 Introduction to Cyber and Information Security Technology IMT4130 Cybercrime Investigation IMT4133 Data Science for Security and Forensics IMT4125 Network Security IMT4016 Eksperter i team - Digital Communities and Welfare IMT4123 Systemsikkerhet IMT4210 Computational Forensics IMT4205 Forprosjekt IIK3100 Etisk hacking og penetrasjonstesting MIS4900 Masteroppgave informasjonssikkerhet

Retningslinjer - rettigheter og plikter

Formål

Avtale om veiledning av masteroppgaven/hovedoppgaven er en samarbeidsavtale mellom student, veileder og institutt. Avtalen regulerer veiledningsforholdet, omfang, art og ansvarsfordeling.

Studieprogrammet og arbeidet med oppgaven er regulert av Universitets- og høgskoleloven, NTNUs studieforskrift og gjeldende studieplan. Informasjon om emnet, som oppgaven inngår i, finner du i emnebeskrivelsen.

Veiledning

Studenten har ansvar for å

- Avtale veiledningstimer med veileder innenfor rammene master-/hovedoppgaveavtalen gir.
- Utarbeide framdriftsplan for arbeidet i samråd med veileder, inkludert veiledningsplan.
- Holde oversikt over antall brukte veiledningstimer sammen med veileder.
- Gi veileder nødvendig skriftlig materiale i rimelig tid før veiledning.
- Holde instituttet og veileder orientert om eventuelle forsinkelser.
- Inkludere eventuell(e) medstudent(er) i avtalen.

Veileder har ansvar for å

- Avklare forventninger om veiledningsforholdet.
- Sørge for at det søkes om eventuelle nødvendige godkjenninger (etikk, personvernhensyn).
- Gi råd om formulering og avgrensning av tema og problemstilling, slik at arbeidet er gjennomførbart innenfor normert eller avtalt studietid.
- Drøfte og vurdere hypoteser og metoder.
- Gi råd vedrørende faglitteratur, kildemateriale, datagrunnlag, dokumentasjon og eventuelt ressursbehov.
- Drøfte framstillingsform (eksempelvis disposisjon og språklig form).
- Drøfte resultater og tolkninger.
- Holde seg orientert om progresjonen i studentens arbeid i henhold til avtalt tids- og arbeidsplan, og følge opp studenten ved behov.
- Sammen med studenten holde oversikt over antall brukte veiledningstimer.

Instituttet har ansvar for å

- Sørge for at avtalen blir inngått.
- Finne og oppnevne veileder(e).
- Inngå avtale med annet institutt/ fakultet/institusjon dersom det er oppnevnt ekstern medveileder.
- I samarbeid med veileder holde oversikt over studentens framdrift, antall brukte veiledningstimer, og følge opp dersom studenten er forsinket i henhold til avtalen.
- Oppnevne ny veileder og sørge for inngåelse av ny avtale dersom:
    - Veileder blir fraværende på grunn av eksempelvis forskningstermin, sykdom, eller reiser.
    - Student eller veileder ber om å få avslutte avtalen fordi en av partene ikke følger den.
    - Andre forhold gjør at partene finner det hensiktsmessig med ny veileder.
- Gi studenten beskjed når veiledningsforholdet opphører.
- Informere veileder(e) om ansvaret for å ivareta forskningsetiske forhold, personvernhensyn og veiledningsetiske forhold.
- Ønsker student, eller veileder, å bli løst fra avtalen må det søkes til instituttet. Instituttet må i et slikt tilfelle oppnevne ny veileder.

Avtaleskjemaet skal godkjennes når retningslinjene er gjennomgått.

Godkjent av

Henrik Hansen Stormyhr
Student

23.01.2025
Digitalt godkjent

Tjerand Silde
Veileder

23.01.2025
Digitalt godkjent

Hilde Bakke
Institutt

05.02.2025
Digitalt godkjent

Master`s Agreement / Main Thesis Agreement

Faculty	**Faculty of Information Technology and Electrical Engineering**
Institute	**Department of Information Security and Communication Technology**
Programme Code	**MIS**
Course Code	**MIS4900**

Personal Information

Surname, First Name	**Stormyhr, Henrik Hansen**
Date of Birth	**30.09.1999**
Email	**henrhst@stud.ntnu.no**

Supervision and Co-authors

Supervisor	**Tjerand Silde**
Co-supervisors (if applicable)	**Emil August Hovd Olaisen, Bor de Kock**
Co-authors (if applicable)	

The Master`s thesis

Starting Date	**06.01.2025**
Submission Deadline	**10.06.2025**
Thesis Working Title	**A comparative analysis of MOS-LQO algorithms for perceptual transparency testing in audio steganography**
Problem Description	**Perceptual transparency testing is highly relevant for evaluating audio steganography methods. The perceptual transparency of such methods says something about how easy it is for a human observer to hear that a secret message has been embedded into an audio file. One way to test this is by using Objective Mean Opinion Score (MOS-LQO) algorithms. These algorithms try to emulate Subjective Mean Opinion Score (MOS-LQS) tests where human participants rate the perceived quality of some altered/degraded audio files. However, there exists a plethora of different MOS-LQO algorithms and no current work is exploring which one is best suited for evaluating the perceptual transparency of different audio steganography methods. This study aims to change this by investigating how some different MOS-LQO algorithms compare to Subjective MOS (MOS-LQS) tests when evaluating the perceptual transparency of some chosen audio steganography tools and methods.**

Risk Assessment and Data Management

Will you conduct a Risk Assessment? **If "Yes", Is the Risk Assessment Conducted?**	**No** **No**
Will you Apply for Data Management? **(REK*, Sikt**)**	**Yes**
Will You Write a Confidentiality Agreement? **If "Yes", Is the Confidentiality Agreement Conducted?**	**No** **No**

* REK -- https://rekportalen.no/

** Sikt's Notification Form for personal data in research (https://sikt.no/en/notification-form-personal-data)

Topics to be included in the Master`s Degree (if applicable)

IMT4110 Scientific Methodology and Communication IMT4115 Introduksjon til informasjonssikkerhetsledelse IMT4114 Introduction to Digital Forensics IMT4113 Introduction to Cyber and Information Security Technology IMT4130 Cybercrime Investigation IMT4133 Data Science for Security and Forensics IMT4125 Network Security IMT4016 Eksperter i team - Digital Communities and Welfare IMT4123 Systemsikkerhet IMT4210 Computational Forensics IMT4205 Forprosjekt IIK3100 Etisk hacking og penetrasjonstesting MIS4900 Masteroppgave informasjonssikkerhet

Guidelines – Rights and Obligations

Purpose

The Master's Agreement/ Main Thesis Agreement is an agreement between the student, supervisor, and department. The agreement regulates supervision conditions, scope, nature, and responsibilities concerning the thesis.

The study programme and the thesis are regulated by the Universities and University Colleges Act, NTNU's study regulations, and the current curriculum for the study programme.

Supervision

The student is responsible for

- Arranging the supervision within the framework provided by the agreement.
- Preparing a plan of progress in cooperation with the supervisor, including a supervision schedule.
- Keeping track of the counselling hours.
- Providing the supervisor with the necessary written material in a timely manner before the supervision.
- Keeping the institute and supervisor informed of any delays.
- Adding fellow student(s) to the agreement, if the thesis has more than one author.

The supervisor is responsible for

- Clarifying expectations and how the supervision should take place.
- Ensuring that any necessary approvals are acquired (REC, ethics, privacy).
- Advising on the demarcation of the topic and the thesis statement to ensure that the work is feasible within agreed upon time frame.
- Discussing and evaluating hypotheses and methods.
- Advising on literature, source material, data, documentation, and resource requirements.
- Discussing the layout of the thesis with the student (disposition, linguistic form, etcetera).
- Discussing the results and the interpretation of them.
- Staying informed about the work progress and assist the student if necessary.
- Together with the student, keeping track of supervision hours spent.

The institute is responsible for

- Ensuring that the agreement is entered into.
- Find and appoint supervisor(s).
- Enter into an agreement with another department / faculty / institution if there is an external co-supervisor.
- In cooperation with the supervisor, keep an overview of the student's progress, the number of supervision hours. spent, and assist if the student is delayed by appointment.
- Appoint a new supervisor and arrange for a new agreement if:
  - The supervisor will be absent due to research term, illness, travel, etcetera.
  - The student or supervisor requests to terminate the agreement due to lack of adherence from either party.
  - Other circumstances where it is appropriate with a new supervisor.
- Notify the student when the agreement terminates.
- Inform supervisors about the responsibility for safeguarding ethical issues, privacy and guidance ethics
- Should the cooperation between student and supervisor become problematic, either party may apply to the department to be freed from the agreement. In such occurrence, the department must appoint a new supervisor

This Master`s agreement must be signed when the guidelines have been reviewed.

Signatures

Henrik Hansen Stormyhr
Student

23.01.2025
Digitally approved

Tjerand Silde
Supervisor

23.01.2025
Digitally approved

Hilde Bakke
Department

05.02.2025
Digitally approved

Appendix B

DMOS Rating Paper

This are the papers given to each participant for each audio steganography method (labeled as "round"). The participants each get five pairs of these papers to give their ratings for the null-pair and our four chosen audio steganography methods and fills out a a total of ten sheets of paper by the end of the test.

Participant age: _____ **Volume:**
Participant sex: Male [ ] | Female [ ] **Participant number:**

DMOS Test - **ROUND ___**

Female Speaker 1 (Speaker 1)

Female Speaker 1 (Speaker 1) – Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Female Speaker 1 (Speaker 1) – Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Female Speaker 2 (Speaker 2)

Female Speaker 2 (Speaker 2) - Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Female Speaker 2 (Speaker 2) - Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Participant age: _____ **Volume:**
Participant sex: Male [ ] | Female [ ] **Participant number:**

DMOS Test - **ROUND ___**

Male speaker 1 (Speaker 3)

Male Speaker 1 (Speaker 3) – Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Male Speaker 1 (Speaker 3) – Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Male Speaker 2 (Speaker 4)

Male Speaker 2 (Speaker 4)

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Male Speaker 2 (Speaker 4)

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

<-*Least annoying (**better** quality)* *Most annoying (**worse** quality)**->***

Appendix C

Consent Form

The consent form used to collect consent from each participant before their participation in the Subjective DMOS test can be seen below. The consent form is only given in Norwegian as this is what was used in the experiment.

Vil du delta i Masteroppgave om «lyd steganografi»?

Formålet med prosjektet
Dette er et spørsmål til deg om du vil delta i et forskningsprosjekt hvor formålet er å undersøke hvor godt vi kan vurdere en metrikk kalt «perceptual transparency» innenfor lydsteganografi feltet algoritmisk.
Lyd steganografi er kunsten i å gjemme en hemmelig melding i en lydfil, slik at den ikke kan oppdages av uvedkommende. «Perceptual transparency» metrikken sier noe om hvor enkelt et menneske kan høre at en lydfil som inneholder en hemmelig melding.

Jeg ønsker derfor å utføre en såkalt «Mean opinion score» test der rundt 20-40 deltakere vil rangere lydkvaliteten til forskjellige lydfiler som både inneholder og ikke inneholder hemmelige meldinger. Hver deltaker vil rangere noen forskjellige lydfiler fra 1-5 basert på opplevd lydkvalitet. Der 1 er dårlig og 5 er glimrende. Gjennomsnittet av rangeringen til alle deltakerne for hver lydfil vil utgjøre hver lydfils «Mean opinion score».

Jeg ønsker også å samle inn ditt kjønn og din alder da deltakerne bør være så balansert som mulig på kjønn og alder for å få et godt resultat.

I forbindelse med min masteroppgave ved NTNU ønsker jeg å sammenlikne resultatene av disse testene med algoritmer som forsøker å emulere de samme testene objektivt. Dette er for å finne ut hvor gode forskjellige algoritmer er til å vurdere «perceptual transparency» metrikken for forskjellige lyd steganografi metoder.

Hvorfor får du spørsmål om å delta?
Du får denne forespørselen fordi jeg trenger deltakere fra forskjellige kjønn og aldersgrupper for å uføre denne testen.

Hvem er ansvarlig for forskningsprosjektet?
Institutt for informasjonssikkerhet og kommunikasjonsteknologi ved Norges teknisk-naturvitenskapelige universitet er ansvarlig for personopplysningene som behandles i prosjektet.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Hva innebærer det for deg å delta?
Varighet ca. 30 minutter.
Jeg vil først spørre deg om kjønn og alder og notere meg dette.

Du vil høre på noen par lydklipp med støydempende hodetelefoner. Etter hvert lydklipp vi du bes om å notere ned opplevd lydkvalitet fra en til fem på et ark.
Jeg vil notere meg disse tallene for hvert lydklipp under ditt kjønn og alder, men jeg vil ikke notere navnet ditt eller noen andre personlige opplysninger.

Kort om personvern
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler personopplysningene konfidensielt og i samsvar med personvernregelverket. Du kan lese mer om personvern på nedenfor.

Med vennlig hilsen

Henrik Hansen Stomyhr (Masterstudent) Tjerand Silde (Veileder)

Utdypende om personvern – hvordan vi oppbevarer og bruker dine opplysninger
De eneste som vil ha tilgang til personopplysningene er meg (Henrik Hansen Stormyhr) og eventuelt mine veiledere for prosjektet ved NTNU (Tjerand Silde, Bor de Kock og Emil August Hovd Olaisen.)

Ditt kjønn og alder vil krypteres med et sikkert passord og lagres på min datamaskin og en ekstern harddisk som backup. Ditt samtykke vil oppbevares på et hemmelig og låst sted og makuleres når prosjektet er over.

De eneste personopplysningene som kan bli publisert er kjønn og alder. Dette vil sannsynligvis bli skrevet noe om med tanke på balansen av deltakeres kjønn og alder.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Norges teknisk-naturvitenskapelige universitet har personverntjenestene ved Sikt – Kunnskapssektorens tjenesteleverandør, vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Dine rettigheter
Så lenge du kan identifiseres i datamaterialet, har du rett til:
- å be om innsyn i hvilke opplysninger vi behandler om deg, og få utlevert en kopi av opplysningene,
- å få rettet opplysninger om deg som er feil eller misvisende,
- å få slettet personopplysninger om deg,
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Vi vil gi deg en begrunnelse hvis vi mener at du ikke kan identifiseres, eller at rettighetene ikke kan utøves.

Hva skjer med personopplysningene dine når forskningsprosjektet avsluttes?
Prosjektet vil etter planen avsluttes innen utløpet av Juni 2025.

Opplysningene vil da slettes fra min datamaskin og ekstern harddisk.
Samtykkeskjemaet vil makuleres.

Spørsmål

Hvis du har spørsmål eller vil utøve dine rettigheter, ta kontakt med:

Henrik Hansen Stormyhr (Masterstudent)
henrhst@stud.ntnu.no
+47 47615188
Eller
Tjerand Silde (Veileder)
tjerand.silde@ntnu.no
+47 47301607
Eller
Thomas Ørnulf Helgesen (Personvernombud ved NTNU)
thomas.helgesen@ntnu.no
+47 93079038
Hvis du har spørsmål knyttet til Sikts vurdering av prosjektet, kan du ta kontakt på e-post: personverntjenester@sikt.no, eller på telefon: 73 98 40 40.

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet Masteroppgave om «lyd steganografi», og har fått anledning til å stille spørsmål. Jeg samtykker til:

- ☐ å delta i «Mean opinion score test» observasjon

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

---

Appendix D

Information paper for participants of the Subjective DMOS test

The information paper given to participants of the Subjective DMOS test is included below. The information paper is only available in Norwegian as this is what was used during the test.

Forklaring av lydkvalitetstest (DMOS DCR ITU-T P.800 Test)

I forbindelse med min masteroppgave i informasjonssikkerhet ved NTNU vil jeg utføre en test for å sammenlikne forverringen av lydkvalitet forårsaket av forskjellige metoder å gjemme hemmelig informasjon i lydfiler. Målet med masteroppgaven er å teste forskjellige metoder og algoritmer for å måle kvaliteten på slike metoder (mer spesifikt hvor hørbare de er for mennesker) og finne svakheter og styrker ved disse metodene. Testen tar ca. 30 minutter og jeg setter veldig pris på alle som kan delta.

Under testen vil deltakerne rangere noen lydfiler fra 1-5 ved bruk av skalaen som kan ses nedenfor:

5	Degradation is inaudible. *Nedgang i lydkvalitet er ikke hørbar.*
4	Degradation is audible but not annoying. *Nedgang i lydkvalitet er hørbar, men ikke irriterende.*
3	Degradation is slightly annoying. *Nedgang i lydkvalitet er litt irriterende.*
2	Degradation is annoying. *Nedgang i lydkvalitet er irriterende.*
1	Degradation is very annoying. *Nedgang i lydkvalitet er veldig irriterende.*

5 ---------- 4 ---------- 3 ---------- 2 ---------- 1

<-Least annoying (better quality) ---------- Most annoying (worse quality)->

<-Minst irriterende (bedre kvalitet) ***Mest irriterende (værre kvalitet) ->***

Rangeringen vil gis på papir til forskjellige «Samples» som består av to setninger som blir gjentatt et par ganger. Avspillingen av disse «samplesene» vil forklares bedre på neste side. Rangeringen vil gis ved avkryssning i et skjema utformet som det som vises nedenfor:

Female Speaker 1 (Speaker 1)

Female Speaker 1 (Speaker 1) – Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 1 (Speaker 1) – Sample 2

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 2 (Speaker 2)

Female Speaker 2 (Speaker 2) - Sample 1

Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

Female Speaker 2 (Speaker 2) - Sample 2

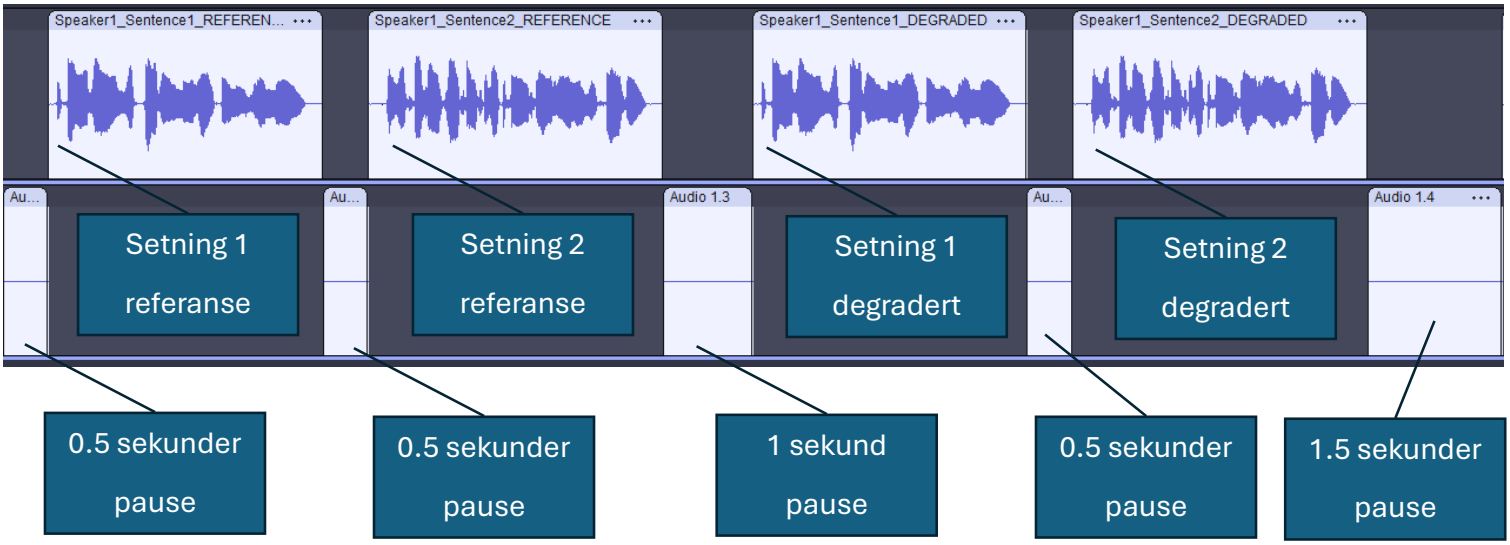
Mark an "X" in the box below your perceived audio quality rating (From 1-5).

5	4	3	2	1

*<-Least annoying (**better** quality)* *Most annoying (**worse** quality)->*

For hver “Sample” på skjemaet vil det spilles lydklipp av to innspilte setnigner lest opp av samme person med 0.5 sekunders mellomrom. Den første gangen disse setningene spilles av vil være referenaselyden. Det vil bli holdt opp et skilt der det står «referense» under avspillingen av disse. Deretter vil det være en pause på 1 sekund før de degraderte lyd-filene spilles av. Dette vil være de samme setningene, men de vil være behandlet på en måte som kan gi varierende grad av degradert/forverret lydkvalitet. Et skilt der det står «Forverret lydkvalitet» vil holdes opp under avspillingen av disse. Denne sekvensen vil gjentas to ganger, slik at deltakeren kan være helt sikkert på om det finnes en degradering/forverring i lydkvaliteten eller ikke. Deretter vil deltaktren rangere «sampelen» fra 1-5 og neste «sample» vil spilles av når jeg ser at rangeringen har blitt gjort. En illustrasjon av dette oppsettet kan sees nedenfor (denne vil gjentas to ganger for hver «sample»):

Illustrasjon av avspillingen av en «sample». Denne sekvensen vil spilles av to ganger:



Degraderingen/forverringen i lydkvalitet kan noen ganger være veldig hørbar, og noen ganger ikke være hørtbar i det hele tatt. Testen går ut på å rangere graden av degradert/forverret lydkvalitet fra 1-5 for hver av disse «samplesene».

Spørsmål vil selvfølgelig bli besvart både før og under testen.

Testen kan høres komplisert ut, men min erfaring så langt er at de fleste forstår oppsettet raskt under selve testen.

Håper du vil delta! 😊

Henrik Hansen Stomyhr

Appendix E

DMOS Degradation scale explanation

The explanation and Norwegian translation of the DMOS degradation scale given to the participants of the Subjective DMOS test is provided below. The sample labels like "F1-S1" in the tables work like this: F or M refers to male or female speaker, the number behind this letter serves to differentiate between the two speakers per gender. The S stands for sample and serves to label the two samples per speaker.

Audio degradation rating scale explanation

Forklaring av rangeringsskala for nedgang i lydkvalitet

5	Degradation is inaudible. *Nedgang i lydkvalitet er ikke hørbar.*
4	Degradation is audible but not annoying. *Nedgang i lydkvalitet er hørbar, men ikke irriterende.*
3	Degradation is slightly annoying. *Nedgang i lydkvalitet er litt irriterende.*
2	Degradation is annoying. *Nedgang i lydkvalitet er irriterende.*
1	Degradation is very annoying. *Nedgang i lydkvalitet er veldig irriterende.*

Scale visualization

Visualisering av skala

5 ---------- 4 ---------- 3 ---------- 2 ---------- 1

<-Least annoying (better quality) ---------- Most annoying (worse quality)->

<-Minst irriterende (bedre kvalitet) ***Mest irriterende (værre kvalitet) ->***

Appendix F

All SNR Results

Our measured SNR results for all samples and methods can be found in the tables below.

Sample	$SNR^{Steghide}$	$DMOS^{Steghide}$	$SNR^{Hide4PGP}$	$DMOS^{Hide4PGP}$
F1-S1	67.2744	4.8421	73.7286	4.6842
F1-S2	66.108	4.6842	72.8344	4.6842
F2-S1	60.8266	4.5790	73.2963	4.4211
F2-S2	70.3024	4.7368	75.6877	4.5790
M1-S1	66.3435	4.8947	70.4023	4.7368
M1-S2	69.6100	5.0000	68.8338	5.0000
M2-S1	70.5566	4.6316	67.2012	4.7368
M2-S2	71.6951	4.6316	68.1834	4.7895
AVG_F	66.1280	4.7105	73.7286	4.5921
AVG_M	69.5513	4.7895	68.6552	4.8158
AVG_{All}	67.8396	4.7500	71.1919	4.7040

Table F.1: Our measured SNR results for all Steghide and Hide4PGP samples, along with the averages for female samples, male samples and all samples.

Sample	SNR^{GANLow}	$DMOS^{GANLow}$	$SNR^{GANHigh}$	$DMOS^{GANHigh}$
F1-S1	41.8081	2.8947	41.8448	2.6842
F1-S2	41.0197	2.6316	41.0461	2.5790
F2-S1	40.4620	2.4211	40.4922	2.2632
F2-S2	40.0506	2.3158	40.0529	2.3158
M1-S1	38.9550	2.9474	39.0194	3.0526
M1-S2	37.3718	2.7368	37.3833	3.0000
M2-S1	35.7744	2.5790	35.6228	2.8421
M2-S2	34.6223	2.3684	34.5545	2.1053
AVG_F	40.8351	2.5658	40.8590	2.4605
AVG_M	36.6809	2.6579	36.6450	2.7500
AVG_{All}	38.7580	2.6118	38.7520	2.6053

Table F.2: Our measured SNR results for all GAN Low and GAN High samples, along with the averages for female samples, male samples and all samples.

Appendix G

All Correlation Scatter Plots

All of the full size correlation scatter plots between DMOS and all MOS-LQO algorithms with regression lines, for all samples, male samples and female samples can be found below.

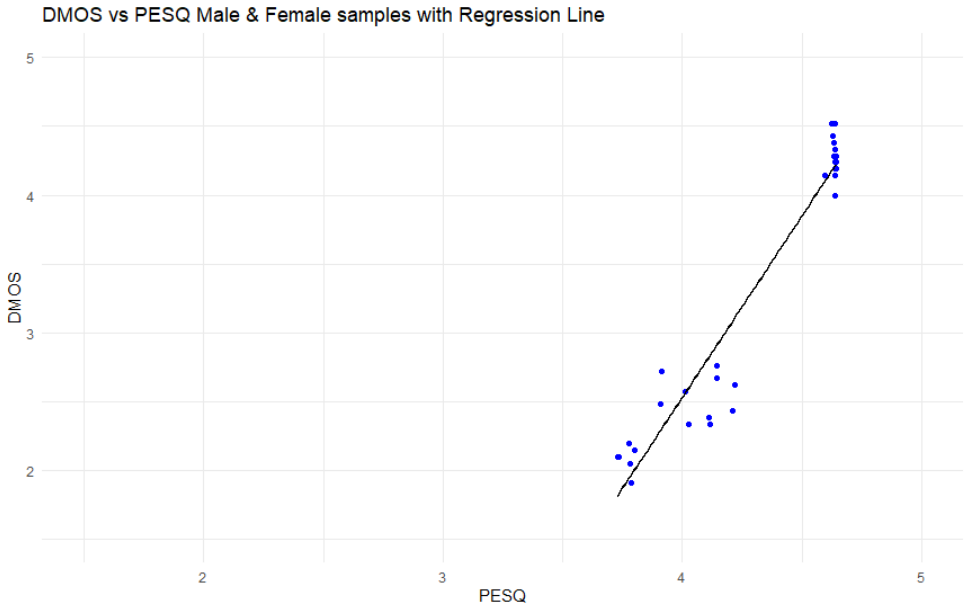


Figure G.1: This scatter plots with a regression line shows the relationship between the $DMOS$ and $PESQ$ scores for all 32 male and female samples.

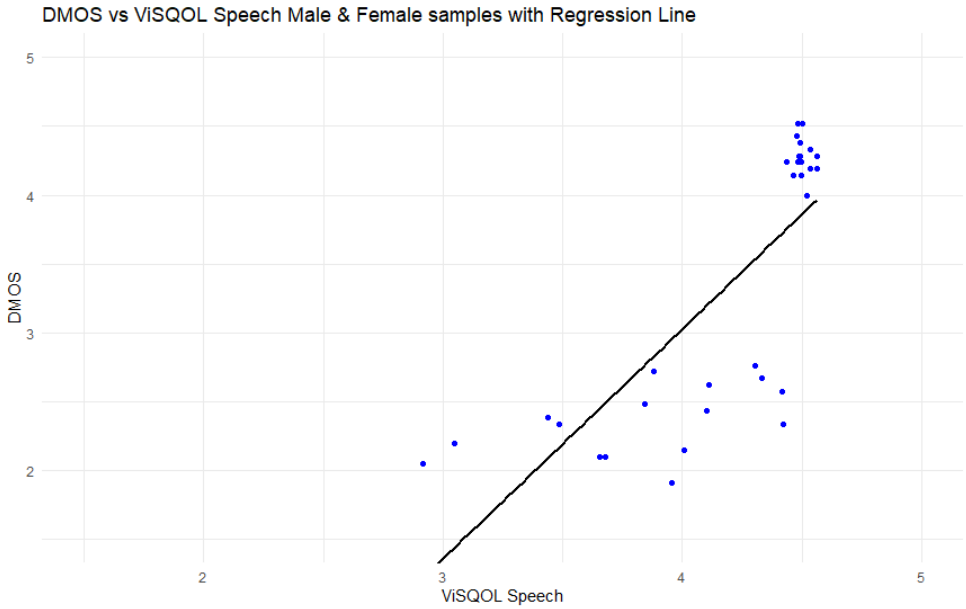


Figure G.2: This scatter plots with a regression line shows the relationship between the $DMOS$ and $ViSQOLSpeech$ scores for all 32 male and female samples.

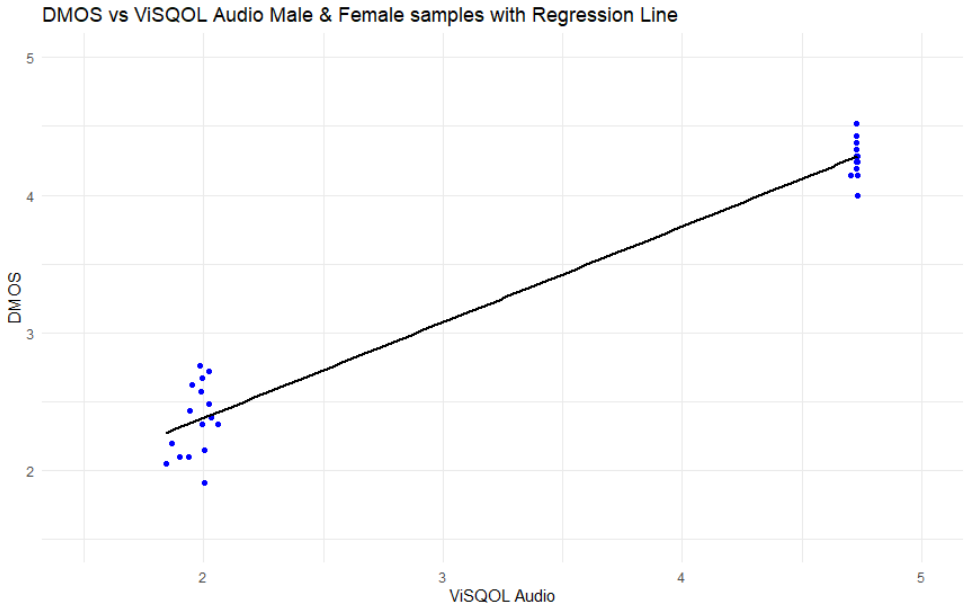


Figure G.3: This scatter plots with a regression line shows the relationship between the $DMOS$ and $ViSQOLAudio$ scores for all 32 male and female samples.

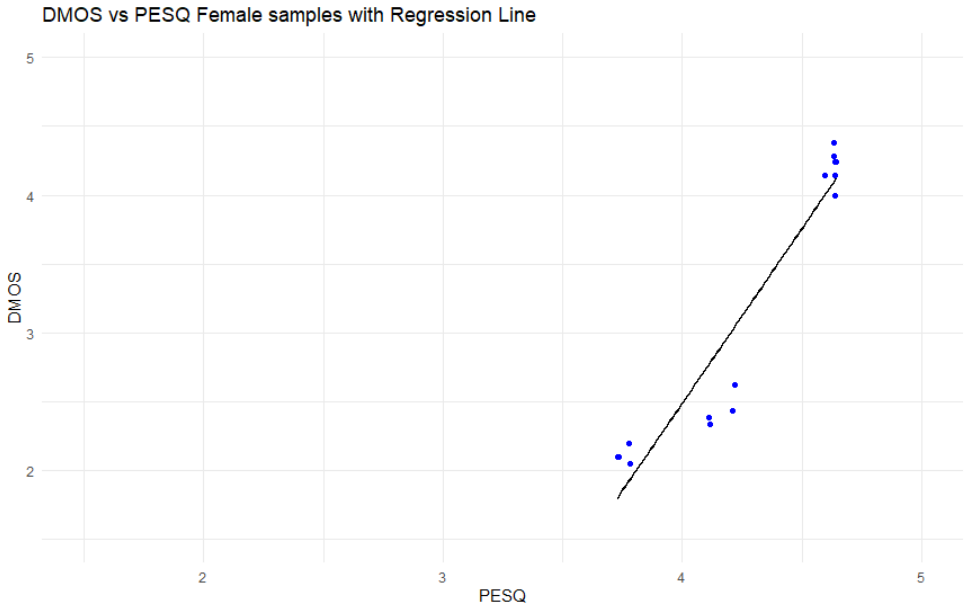


Figure G.4: This scatter plots with a regression line shows the relationship between the $DMOS_F$ and $PESQ_F$ scores for the 16 female samples.

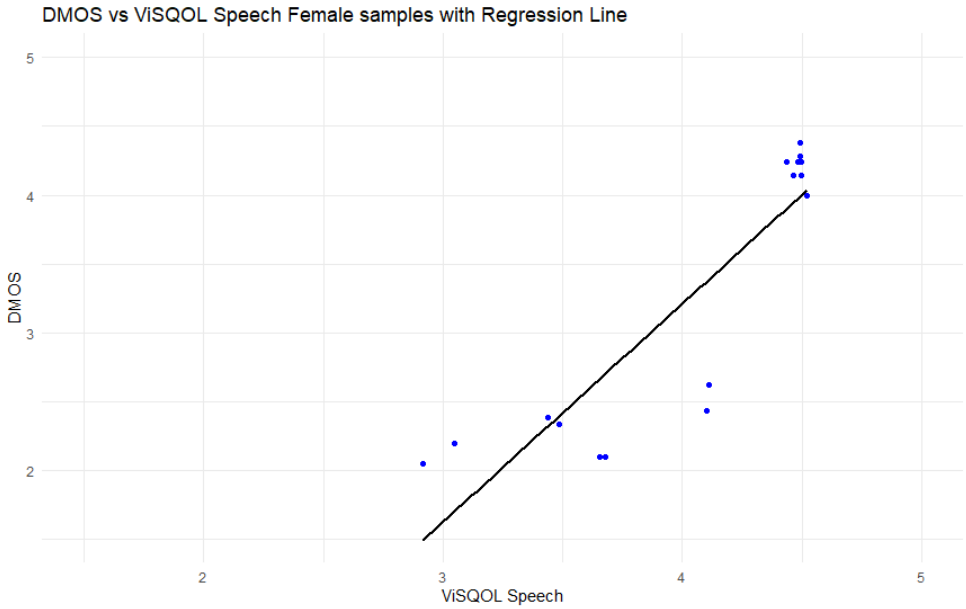


Figure G.5: This scatter plots with a regression line shows the relationship between the $DMOS_F$ and $ViSQOLSpeech_F$ scores for the 16 female samples.

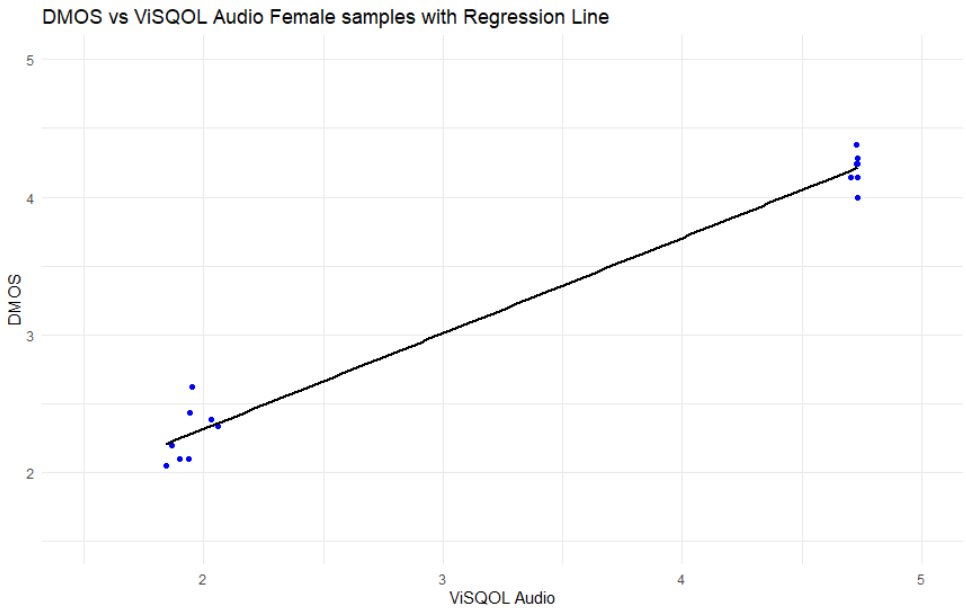


Figure G.6: This scatter plots with a regression line shows the relationship between the $DMOS_F$ and $ViSQOLAudio_F$ scores for the 16 female samples.

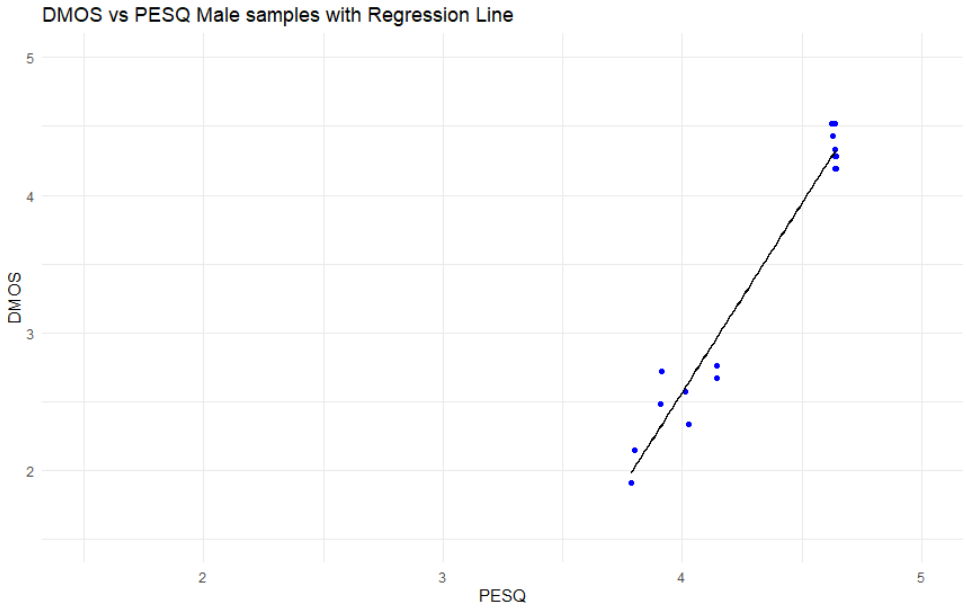


Figure G.7: This scatter plots with a regression line shows the relationship between the $DMOS_M$ and $PESQ_M$ scores for the 16 male samples.

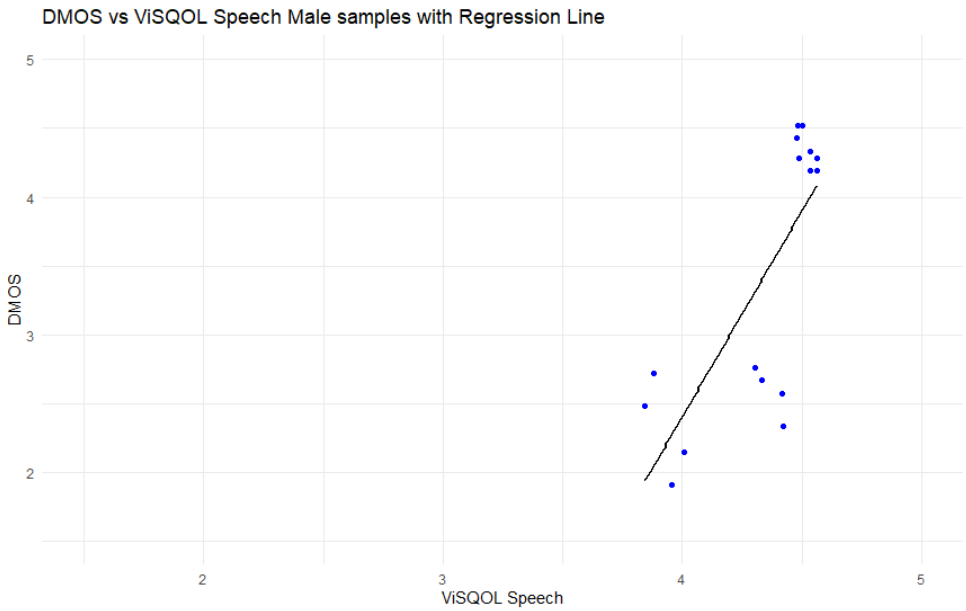


Figure G.8: This scatter plots with a regression line shows the relationship between the $DMOS_M$ and $ViSQOLSpeech_M$ scores for the 16 male samples.

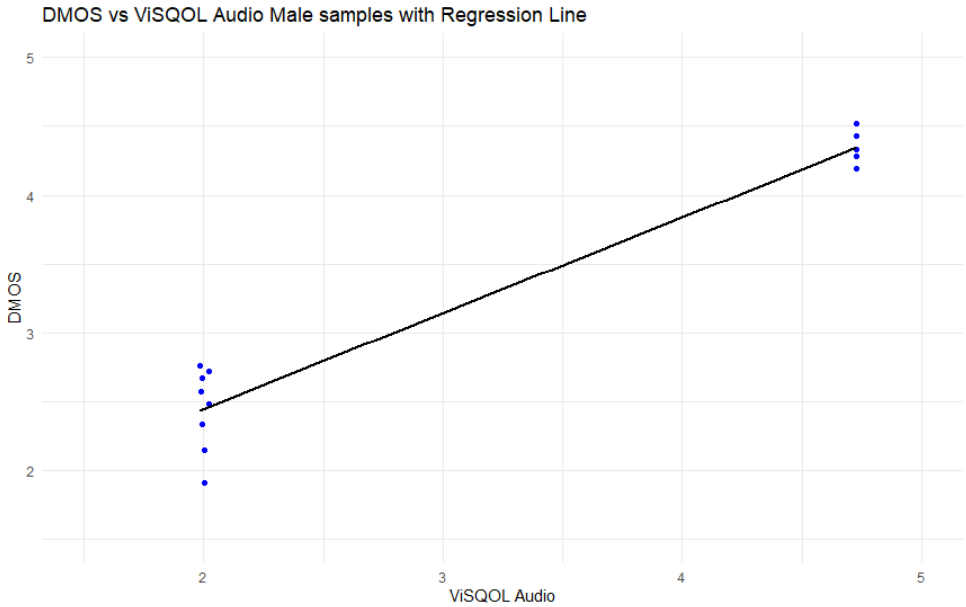
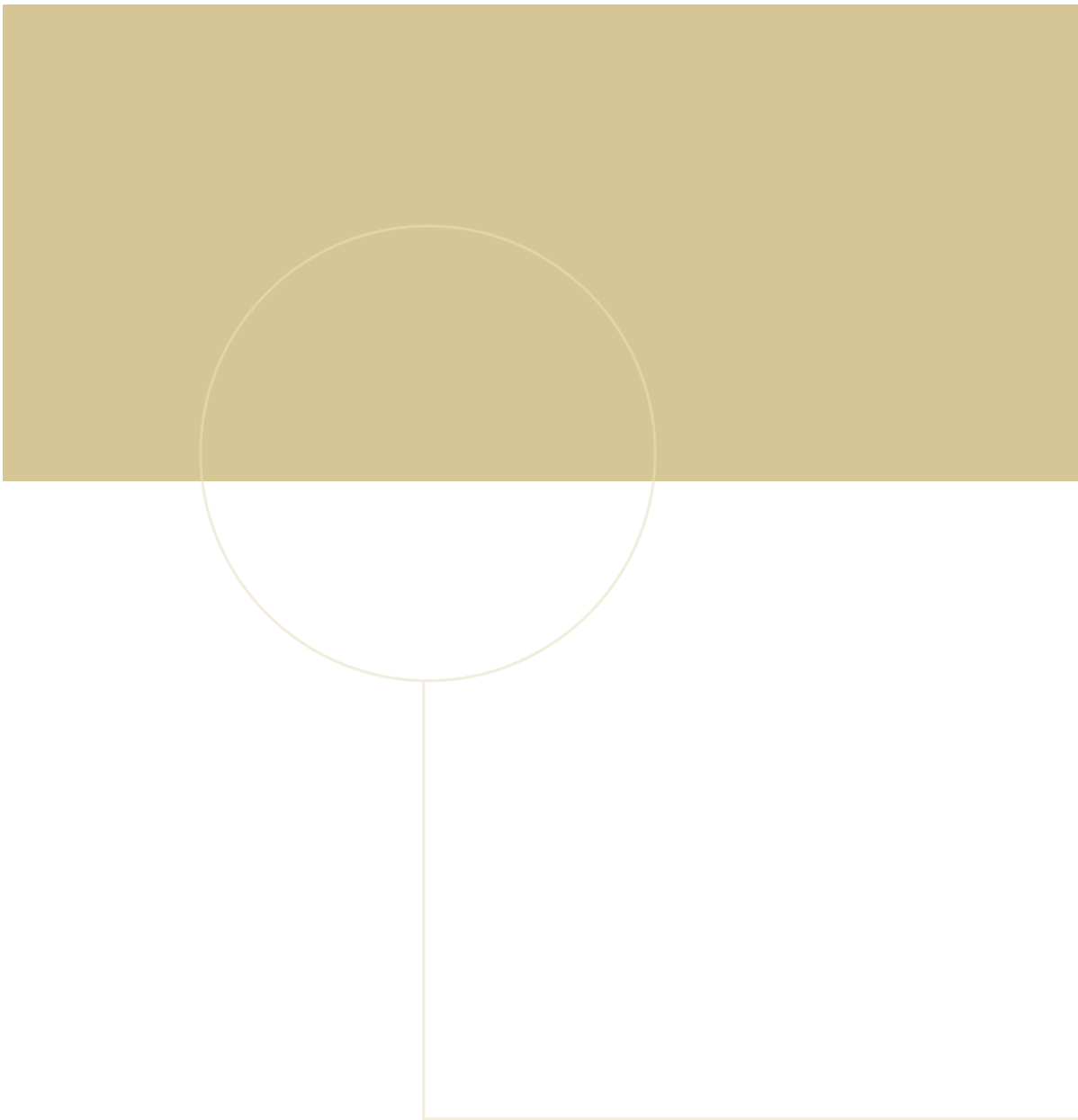


Figure G.9: This scatter plots with a regression line shows the relationship between the $DMOS_M$ and $ViSQOLAudio_M$ scores for the 16 male samples.



NTNU

Norwegian University of
Science and Technology