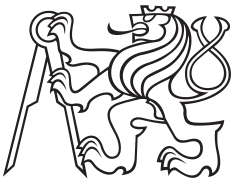
Bachelor Project



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Measurement

Watermarked speech subjective testing

Lucie Mühlfeitová

Supervisor: prof. Ing. Jan Holub, Ph.D.
Field of study: Cybernetics and Robotics
August 2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Mühlfeitová** Jméno: **Lucie** Osobní číslo: **483665**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra měření**

Studijní program: **Kybernetika a robotika**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Subjektivní testování kvality nahrávky řeči s vodoznakem

Název bakalářské práce anglicky:

Watermarked speech subjective testing

Pokyny pro vypracování:

Na základě rešerše existujících mezinárodních doporučení pro subjektivní testování kvality přenosu hlasu a současných metod vodoznakování řeči navrhněte a realizujte modelový subjektivní test, např. dle metodiky P.800 DCR, případně MUSHRA. Posuďte vhodnost využití paralelní úlohy. Na základě statistického vyhodnocení testu porovnejte posuzované algoritmy vodoznakování a navrhněte vhodné oblasti využití.

Seznam doporučené literatury:

[1] ITU-T P.800
[2] ETSI ETR 103 503
[3] Hofbauer, K: Speech Watermarking and Air Traffic Control, Faculty of Electrical and Information Engineering, Graz University of Technology, Austria

Jméno a pracoviště vedoucí(ho) bakalářské práce:

prof. Ing. Jan Holub, Ph.D., katedra měření FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **25.01.2021** Termín odevzdání bakalářské práce: **13.08.2021**

Platnost zadání bakalářské práce:
do konce zimního semestru 2022/2023

prof. Ing. Jan Holub, Ph.D. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Studentka bere na vědomí, že je povinna vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. | Podpis studentky

Datum převzetí zadání

Acknowledgements

I would like to thank my supervisor prof. Ing. Jan Holub Ph.D. for patience and help during working on this thesis. And I would also thank my family for supporting me.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

V Praze, ...

Abstract

The aim of this bachelors theses is, based on the study of current watermarking methods and the recommendations for subjective testing ITU-T P.800 [4], to design and implement subjective testing experiment of the quality of the watermarked speech recording. The goal was to find the border value of the watermark strength at which the quality of the recording is not yet affected. The subjective test was performed using the ACR method. The results were analysed, plotted on graphs, statistical t-test was performed and the results were compared with the results of a similar test performed previously.

Keywords: digital watermark, robustness, subjective testing, Absolute Category Rating, ITU-T P.800

Supervisor: prof. Ing. Jan Holub, Ph.D.

Abstrakt

Cílem této bakalářské práce je, na základě prostudování současných metod vodoznakování a normy pro subjektivní testování ITU-T P.800 [4], navrhnout a zrealizovat subjektivní test kvality nahrávky řeči s vodoznakem. Cílem bylo najít hraniční hodnotu síly vodoznaku, při které ještě není ovlivněna kvalita nahrávky. Subjektivní test byl proveden metodou ACR. Výsledky byly analyzovány, vyneseny do grafů, byl proveden statistický t-test a výsledky byly srovnány s výsledky obdobného testu provedeného dříve.

Klíčová slova: digitální vodoznak, robustnost, subjektivní testování, Absolute Category Rating, ITU-T P.800

Překlad názvu: Subjektivní testování kvality nahrávky řeči s vodoznakem

Contents

Figures

Tables

Chapter 1

Introduction

The aim of my thesis is to design and conduct a subjective testing experiment focusing on the impact of the speech watermark.

In the fist part of my thesis I summarise basic information about digital watermark with the focus on speech watermark. In the next chapter the basics of the subjective testing according to the ITU-T P.800 recommendations [4] are given. In the next chapters I describe the process of designing and conducting the experiment. And in the final parts I analyse the outcomes and draw conclusions.

1.1 Motivation

The speech digital watermark is still less researched than other types of watermarks. However it is a promising technology that might find a lot of utilisation in the future. It has been already used in the air traffic control [3] and other fields of utility will most likely follow. There is a need for further research in this area and that is why I chose this topic for my bachelors theses.

Chapter 2

Digital Watermarking

Rapid development in communication technology and also devices that can duplicate and change the content led to a need for an algorithm to protect the property rights of the media. One of the ways how to secure the content is by using digital watermarking.

Digital watermarking lies in inserting extra information into the original file. Ideally, such a watermark should be imperceptible and should be difficult to remove without altering the original content. [9]

There are three main requirements concerning digital watermarking: capacity, robustness and imperceptibility. Capacity states the number of bits of a watermark that can be embedded into the original file. Robustness provides resistance against intentional and unintentional alterations of the media. Imperceptibility is a property that defines detectability of the watermark. These three requirements conflict with each other so they need to be balanced according to specific usage demands. [8]

The aim of the subjective test is to find the limits when the watermark is as robust as possible yet still remains imperceptible.

Digital watermarking can be classified based on several different factors. Based on robustness we distinguish robust, semi-fragile and fragile. Robust digital watermarking detects the watermark even under serious manipulation. Semi-fragile digital watermarking detects the watermark if only small unintentional manipulation was made. Fragile digital watermarking detects the watermark only if there was no manipulation at all.

According to a different field of application, we distinguish three main categories: signal watermarking (audio, speech), multimedia watermarking (image, video) and document watermarking (text, software). [8]

In this work, I focus on speech watermark.

2.1 Speech watermark

Digital speech watermark process is shown in Figure 2.1.

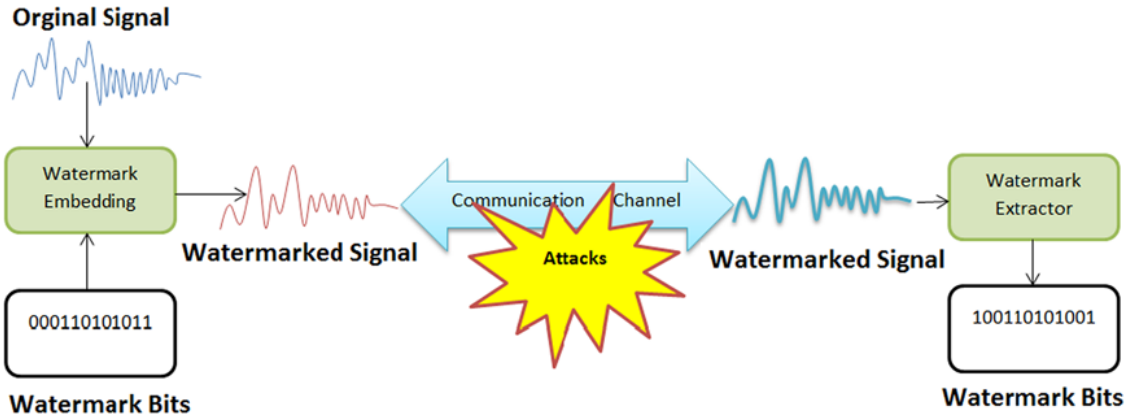


Figure 2.1: Fundamental architecture of digital speech watermarking. [7]

Similarly to digital watermarking in general, we can classify digital speech watermark according to different criteria. In terms of robustness we have robust and fragile digital speech watermarking. In digital speech watermarking, robustness is easier to obtain than fragility. [7]

According to the source and extraction module for digital speech watermarking, we speak about three main categories. Blind speech watermarking which does not need any extra information such as original signal or watermark bits for extraction. Semi-blind speech watermarking which needs extra information like access to the published watermark signal. Non-blind speech watermarking that needs both the original signal and the watermarked signal. [7]

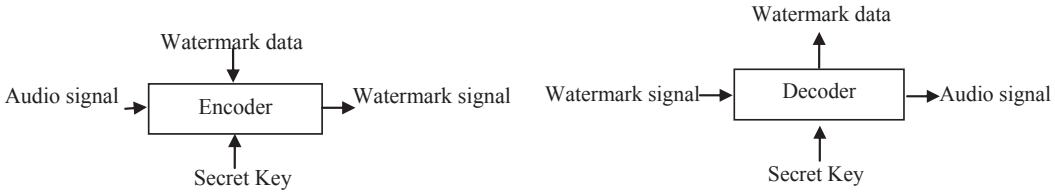
2.2 Embedding digital audio watermark

Current audio watermarking methods can be generally put into two main categories, time domain and transform domain methods. Time domain methods can be further divided into time aligned and echo-based methods. Transform domain methods incorporate spread spectrum, quantization index modulation, and patchwork methods. [6]

In my experiment, an open-source software called Audiowmark developed by Stefan Westfeld was used to embed the watermark into our speech samples. Audiowmark reads the sound file and stores a 128-bit message in the output sound file. The 128-bit message can be later retrieved from the sound file.

As Audiowmark is open-source software, we must ensure that the watermark bits will not be retrieved by an unauthorised user. For that purpose, we use a secret watermarking key.

Then the watermark cannot be retrieved without the right key. A simple diagram of the embedding and retrieving process is shown in Figure 2.2



(a) : Embedding process of watermark **(b) :** Recovery process of watermark

Figure 2.2: Embedding and recovery of watermark. [1]

The software uses blind decoding so the original audio file is not needed to retrieve the message. Audiowmark enables the user to insert the watermark in various strengths. The stronger the watermark, the more robust against different modifications. However, the stronger the watermark, the more audible it becomes. This property is essential for my experiment where I try to find the boundary strength where the watermark still does not affect the subjective quality of the watermarked file.

Audiowmark uses a patchwork algorithm to hide the data in the spectrum of the sound file. The signal is split into 1024 sample frames where for each sample frame, pseudo-randomly selected amplitudes of the frequency bands of a 1024-value fast Fourier transforms are slightly increased or decreased. The used algorithm is inspired by [12].

Chapter 3

Subjective testing

Generally, when dealing with the measuring transmission quality of a signal, we use two main methods. They are objective testing and subjective testing. In general, objective testing is less demanding, cheaper and less time-consuming. Nevertheless, there are situations when it cannot be used. Namely when testing a new technology where there is no objective testing algorithm available. Furthermore, objective testing tends to be less accurate. On the other hand, subjective testing is generally more accurate and can be used for almost every technology.

Depending on the tested technology, different methods of subjective tests are used. Conversation-opinion tests are used when there is an interaction between communicators. For example when testing delay, echo, etc. The most widely used testing methodology and also the methodology I use in my experiment is the listening-opinion test. [4]

3.1 Listening-opinion tests

Commonly used methods in listening-opinion tests are Absolute Category Rating (ACR) method, Degradation Category Rating (DCR) method and Comparison Category Rating (CCR) method. ACR method is the most commonly used one. [4]

DCR method is a modified version of the ACR method and enables higher sensitivity in distinguishing among good quality samples. In this method, the subjects listen to a couple of samples where the first sample is a high-quality reference and the second one is the examined sample. For rating, the subjects use a five point degradation category scale where 5 is the best score meaning that degradation is inaudible and 1 is the lowest grade meaning that degradation is very annoying. [4]

CCR method is very similar DCR method. The difference between these two methods is that while in the DCR method the reference sample is always played first, in CCR the order of the samples is chosen randomly. The grading scale used in this method ranges from 3 to -3

where 3 means that the second sample is much better than the first one and -3 means that the second sample is much worse. [4]

After the consultation with my supervisor we decided to use the Absolute Category Rating method instead of the Degradation Category Rating method that we have originally planned to use. We decided to use the ACR method because it is compatible with objective testing algorithms and therefore more suitable for future steps.

3.1.1 Absolute Category Rating

Sample preparation

The samples should consist of short simple sentences that are easy to understand. The sentences should target common topics without any complicated or technical terms. The sentences should be chosen randomly so there is no evident connection among them. [4]

The samples should be recorded by at least four talkers, two men and two women. The female voices and also the male voices should have a different pitch, one lower and one higher. As individual technologies impact different voices differently, it is important that the samples contain various types of voices. [4]

Every test should include a high quality reference condition. Other conditions used in the experiment are made according to the test purpose. [4]

Listening test procedure

The subjects who are taking part in the listening experiment should be from different age groups with a balanced number of males and females. A bigger variety in testing subjects is important for the validity of the experiment as the hearing ability differs depending on age and gender. Only subjects who have not been taking part in any listening-opinion test for at least a year are allowed to participate in the test. [4]

Previous to the testing, the subjects should be given clear written instructions. They clearly explain the testing procedure and the rating method. The samples are presented to the subjects in random order which is different for every listener. The subjects rate each sample according to a listening-quality scale presented in Table 3.1. [4]

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 3.1: Listening-quality scale. [4]

Analysis of the results

The results of the test are presented as a mean opinion score (MOS). Mean opinion score is the arithmetic mean over all values belonging to one condition.

$$MOS = \frac{1}{L \cdot T} \cdot \sum_{i=1}^{T} \sum_{j=1}^{L} X_{c,i,j}, \tag{3.1}$$

where L is the number of listeners who rated the condition, T is the number of samples belonging to the examined condition and $X_{c,i,j}$ is the score given to the sample i by subject j.

Chapter 4

The experiment

4.1 General considerations of experiment

My experiment was conducted on the 8th and 9th of April 2021. All subjects were Chinese so English, which was the language of the speech samples, was not their native language. The average age of the participants was 34.6 years with standard deviation 12.6 years.

In my experiment 16 conditions were tested. Every condition was represented by 12 different samples so all together we had 192 samples. Twelve listeners participated in our experiment. Each subject heard and rated every sample.

4.2 Samples

For this experiment speech samples published by ETSI were used to ensure the correctness of the samples. The chosen samples were recorded by both, male and female speakers. We chose twelve reference samples, six recorded by male speakers and six by female speakers. Then we adjusted the samples using the Adobe Audition so they fit the requirements of the experiment. The final samples were each exactly 4 seconds long.

The purpose of the experiment was to determine impact of a watermark on the perceivable quality of the audio recording. To embed the watermark into the samples we used an open-source software called Audiowmark as mentioned in Chapter 2. Audiowmark enables embedding the watermark of various strengths which is essential for my experiment. In my experiment I used watermarks with strengths ranging from 10 to 650 and I tried to find the highest possible strength value where the watermark still does not reduce the perceivable quality of the audio file.

Some background noise was added into several samples. The purpose of background noise was to determine the impact of the watermark on recordings taken in real-life environment.

In my experiment, I used two types of background noises. The first background noise used is a simulated noise of the engine from HMMWV tactical transport with a 3 dB Signal to Noise Ratio (SNR). The second one is a simulated pub noise with a 6 dB Signal to Noise Ratio.

Besides the original studio recordings, acoustic recordings with some mild effects such as reverb added were used.

Different watermarking strengths were embedded into different samples. Watermarks with strength 10, 30, 75, 200 and 650 were embedded into the original studio recording. The samples with added pub noise or engine noise from HMMWV tactical transport were watermarked with strengths 10 and 30. And the acoustic recordings were watermarked with watermarks with strengths of 30, 100 and 500.

We chose the distribution of the watermarks and their strengths and also the background noise and its Signal to Noise Ratio by expert listening.

The 16 conditions used in the experiment are described in Table 4.1.

Condition	Studio/Acoustic recording	Background noise	Watermark strength
C01	Studio	-	-
C02	Studio	-	10
C03	Acoustic	-	-
C04	Studio	-	30
C05	Studio	Pub noise	-
C06	Studio	Pub noise	10
C07	Studio	-	75
C08	Studio	HMMWV tactical vehicle noise	10
C09	Studio	HMMWV tactical vehicle noise	-
C10	Studio	Pub noise	30
C11	Studio	-	200
C12	Studio	HMMWV tactical vehicle noise	30
C13	Acoustic	Pub noise	30
C14	Studio	-	650
C15	Acoustic	-	500
C16	Acoustic	-	100

Table 4.1: Conditions used in the experiment.

Chapter 5

T-test

Student's t-test is a frequently used statistical test. It can be used to determine whether two means are different with a given probability of 1-p. [10]

In general, there are three types of Student's t-tests, one-sample t-test, two-sample t-test and paired t-test.

In one-sample t-tests, we compare a single mean with a fixed value. Two-sample t-test, also known as an independent samples t-test is the most commonly used one. It is used to compare the means of the different sets of data. Paired t-test, also known as dependent samples t-test is also used to compare the means of two sets of data. The difference between a two-sample t-test and a paired samples t-test is that the samples in a paired t-test have to be somehow related. They might for example be data from the same people before and after some practice. [2]

In my experiment, I used the dependent samples t-test.

5.1 Performing a paired t-test

The first step is to state a null hypothesis. It assumes that the means are equal ($H_0 : \mu_1 = \mu_2$). Or more precisely that the pairwise difference between the sample data equals zero ($H_0 : \mu_d = 0$).

Then we state an alternative hypothesis (H_1), such as one of the means is higher than the other or that they are just different. We assume that the null hypothesis is true.

The process of making a paired t-test is the same as making a one-sample t-test. First we have to get that one set of data from the two that we have. If we label the first set of data 'X' and the second set of data 'Y' then we would get our desired data by pairwise subtracting

the 'Y' data from 'X'. We will label our new data as 'D'.

$$d_i = y_i - x_i. \tag{5.1}$$

Now we have a single sample set of difference scores. Now we can run a one-sample t-test with the data we have.

A t-value is computed as

$$t = \frac{\overline{d} - (\mu_y - \mu_x)}{\frac{s_d}{\sqrt{n}}}, \tag{5.2}$$

where $\overline{d}$ is the arithmetic mean of the set of difference scores, μ_x and μ_y are the means of the first and second set of data, s_d is the standard deviation of the set of difference scores and n is the number of samples in our data set. [2]

The degrees of freedom is the number of samples subtracted by one ($df = n - 1$).

5.2 Analysing the results

First we need to set a parameter α. It is a significance criterion. In my tests I use $\alpha = 0.05$. That means that if there is a 5% chance or bigger that the difference occurred by accident then the difference is not considered to be statistically significant.

There is a given a critical t-value for each combination of degrees of freedom (df) and significance criterion (α). The critical t-value is a border between a statistically significant difference and a difference which may have occurred by chance. The critical t-value can be found in a specific t-table or it can be computed in Microsoft Excel with a function TINV. If our computed t-value is larger than the critical t-value then we say that the difference is statistically significant at α.

Chapter 6

Data Analysis

6.1 Studio recording with increasing watermarking strength

First, we study the impact of the watermark on the clean high-quality studio recorded samples.

Condition	MOS value
C01 - reference condition	4.86
C02 - watermark strength 10	4.90
C04 - watermark strength 30	4.77
C07 - watermark strength 75	3.92
C11 - watermark strength 200	2.96
C14 - watermark strength 650	1.38

Table 6.1: MOS values of the watermarked conditions.

We compare MOS values of a reference clean studio recording and watermarked samples of increasing strength without any background noise. In Figure 6.1 we can see that the speech quality is decreasing with increasing watermark strength. However, the sensible deterioration does not start until the watermark strength 75. The MOS value of the condition with watermark strength 10 is even higher than the MOS value of the clean reference.

Besides MOS values there are 95% confidence intervals of both the reference and the watermark condition plotted on the graph. An easy way to tell whether there is a statistically significant difference between the MOS values of the conditions is to see if the confidence intervals of the two conditions overlap or not. If they overlap, the difference is most likely not statistically significant. If they do not overlap, the difference is probably statistically significant. By looking at the graph we can tell that the samples with watermarking strengths 75, 200 and 650 are statistically significantly worse than the clean reference while the difference between the samples with watermarking strengths 10 and 30 and the reference sample is not statistically significant.

To prove this hypothesis, a paired samples t-test was performed. The results of the t-test

are presented in Table 6.2. The critical value is 1.98. If the absolute value of the t-value is larger than the critical value then the difference between the samples is statistically significant. If it is lower then the difference is not statistically significant. Positive t-value means that the results of the examined condition are worse than the reference condition. A negative t-value says that the examined condition was rated better than the reference condition. Statistically significant differences are marked with *.

The studio recorded reference sample was rated by the listeners with a MOS value 4.86. All MOS values of the conditions examined in 6.1 are presented in Table 6.1.

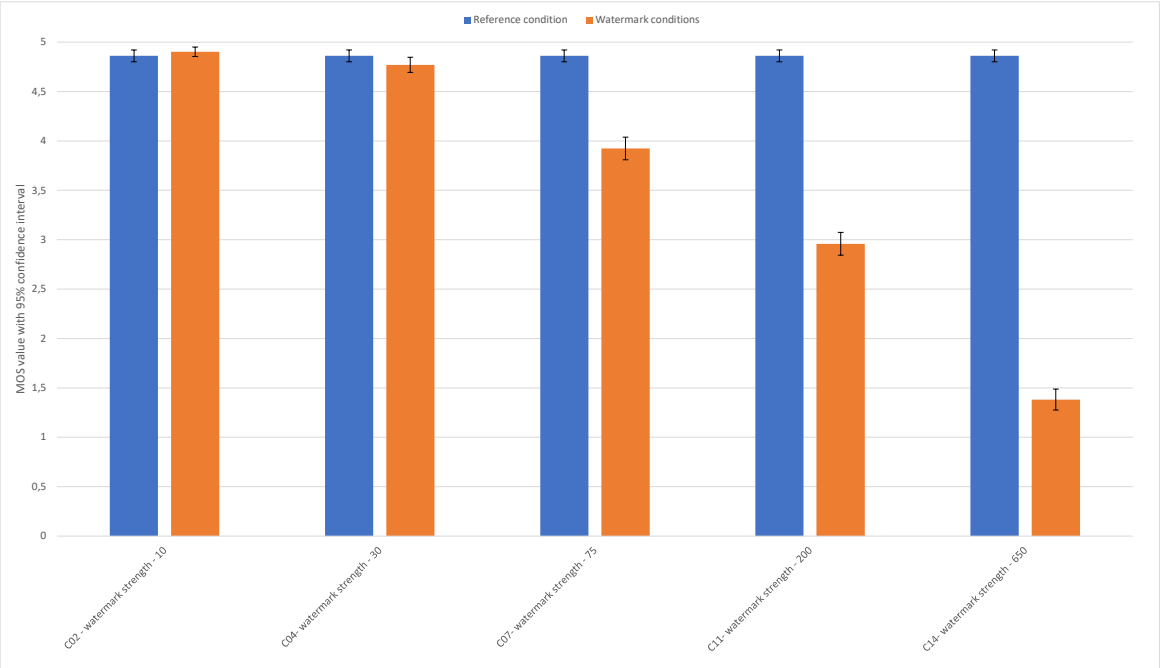


Figure 6.1: Reference studio condition with increasing watermark strength.

Condition	Reference Condition	t-value
C02 - watermark strength 10	C01	-1.23
C04 - watermark strength 30	C01	1.91
C07 - watermark strength 75	C01	*13.99
C11 - watermark strength 200	C01	*27.78
C14 - watermark strength 650	C01	*54.49

Table 6.2: Results of the t-Test performed on the studio conditions with increasing watermarking strength.

6.2 Noise conditions with increasing watermark strength

First, we tested the impact of the background noise on the quality of the recording. MOS values of the noise conditions are presented in Table 6.3.

Condition	MOS value
C01 - reference condition	4.86
C03 - acoustic recording	3.73
C05 - pub noise 30	3.23
C09 - HMMWV noise	3.47

Table 6.3: MOS values of the conditions with background noise.

In Figure 6.2, MOS value of the clean reference sample and MOS values of samples with background noise are compared. We can see that the background noise quite significantly decreases the quality of the speech sample.

Based on the 95% confidence intervals, we can assume that the quality difference between the clean studio recording and the samples with added background noise or the acoustically recorded samples is statistically significant, therefore, the difference between the MOS values does not happen by coincidence.

As expected, the added background noise decreases the perceptional quality of the recording. Though the MOS values of the noise conditions are still somewhere between 3 and 4, as the listeners mostly rated the samples as good or fair.

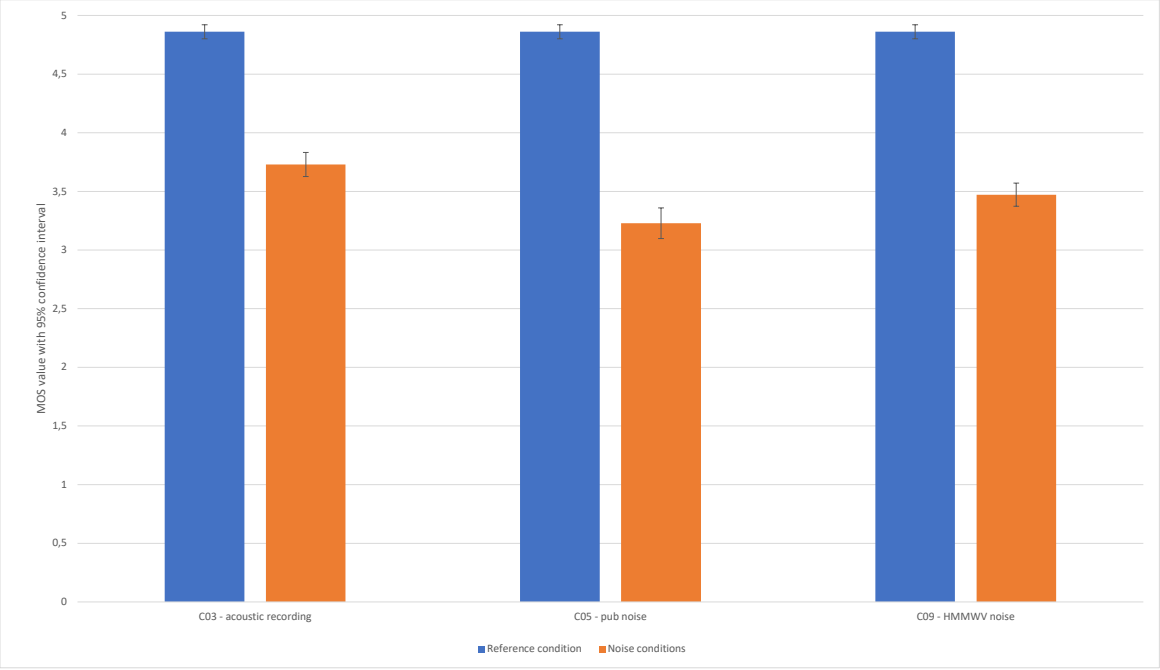


Figure 6.2: Reference studio condition and conditions with background noise without watermark.

The next step is to determine the impact of the watermark on these noise conditions. The knowledge of the impact of the watermark on the recordings with background noise is important because these conditions better simulate the real-life situations and the impact of the watermark might be different than on the high quality studio recordings.

In this case, I do not compare the examined conditions with the high quality studio reference because I have already found out that the background noise statistically significantly reduces the quality of the recording, and I do not expect the watermark to increase the quality, so it is unnecessary to compare it. Instead of that, we compare the examined conditions with the conditions with background noise. We always only compare conditions with the same kind of distortion (pub noise, HMMWV tactical vehicle noise, acoustic recording) with and without watermark.

The MOS values of all the conditions examined here are presented in Table 6.4.

Condition	MOS value
C05 - pub noise without watermark	3.23
C06 - pub noise, watermark strength 10	3.31
C10 - pub noise, watermark strength 30	3.28
C09 - HMMWV noise without watermark	3.47
C08 - HMMWV noise, watermark strength 10	3.58
C12 - HMMWV noise, watermark strength 30	3.46
C03 - Acoustic recording without watermark	3.73
C16 - Acoustic recording, watermark strength 100	3.76
C15 - Acoustic recording, watermark strength 500	2.10

Table 6.4: MOS values of the watermarked conditions with background noise.

The comparisons of the MOS values are plotted in Figure 6.3.

Based on the 95% confidence intervals we can see that the only condition where the watermark statistically significantly decreases the quality of the speech sample is condition 15 which is an acoustic recording watermarked with watermarking strength 500. The quality of other samples seem not affected by the watermark at all. Some watermarked conditions were even rated with higher MOS value than their reference conditions.

To prove the hypotheses I claimed based on analysing the graph with the confidence intervals, I again performed the dependent samples t-test. The results of the t-test with the t-values are presented in Table 6.5. The critical value is again 1.98.

We can see that the t-values of conditions C06, C10, C08 and C16 are negative. That means, as stated before, that they were rated with higher MOS values than their non watermarked reference. However, any of these differences are not statistically significant so we can assume that it most likely only happened by chance. The only statistically significant difference in the quality appeared in condition C15 which is watermark with strength 500 embedded into the acoustic recording. That is really strong watermark so the deterioration of the quality was expected.

Condition	Reference Condition	t-value
C06 - pub noise, watermark strength 10	C05	-1.08
C10 - pub noise, watermark strength 30	C05	-0.93
C08 - HMMWV noise, watermark strength 10	C09	-1.90
C12 - HMMWV noise, watermark strength 30	C09	0.24
C16 - Acoustic recording, watermark strength 100	C03	-0.46
C15 - Acoustic recording, watermark strength 500	C03	*20.42

Table 6.5: Results of the t-Test performed on the noise conditions with increasing watermarking strength.

An interesting thing about the results is that while the watermark with strength 75 embedded into the high quality studio sample caused a statistically significant reduction of the perceptual quality, the watermark with the strength 100 embedded into the acoustically recorded sample did not.

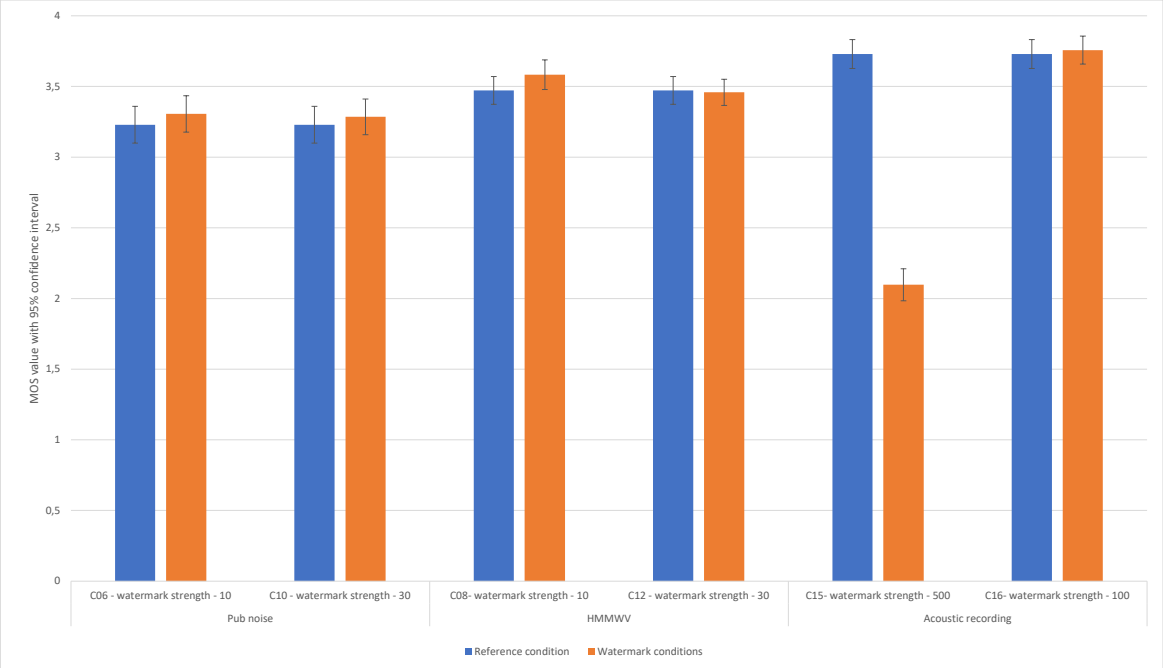


Figure 6.3: Reference noise conditions and noise conditions with increasing watermarking strength.

6.3 Comparison with the previous experiment

Before my experiment, a smaller test with the same samples was conducted. [5]. Besides the number of subjects, the difference between the experiments was the nationality of the listeners. While in my experiment the listeners were Chinese, the listeners in the previous experiment were all either native English speakers or people with excellent knowledge of English.

The comparison of the MOS values from these experiments was plotted on the graph in Figure 6.4. The horizontal axis represents the MOS values from the previous smaller experiment, referred to as the first experiment. The vertical axis shows the MOS values from my experiment, referred to as the second experiment. Again, besides the MOS values, the 95% confidence intervals are also plotted on the graph. There is also a black dotted line plotted on the graph. This line represents the balance between the two tests. If a test condition was rated with the same MOS value in both experiments, it would lay on this line.

If the condition was rated better in the first test than in the second, it would appear below the black dotted line. If the condition was rated better in the second test, it would appear above the black line. We can see that all the blue dots that are representing the MOS values of the conditions from both tests are located above the black dotted line. That means that all of the conditions were rated with higher MOS values in my experiment than in the first one.

If we check the confidence intervals, we can see that none of the intervals crosses the line and only one condition has a confidence interval that touches the line. This information tells us that the differences between the MOS values are statistically significant and most likely did not happen by chance. The condition whose confidence interval touches the line is C05, which is a non watermarked condition with added pub noise. The difference of the MOS values of this condition might be a coincidence.

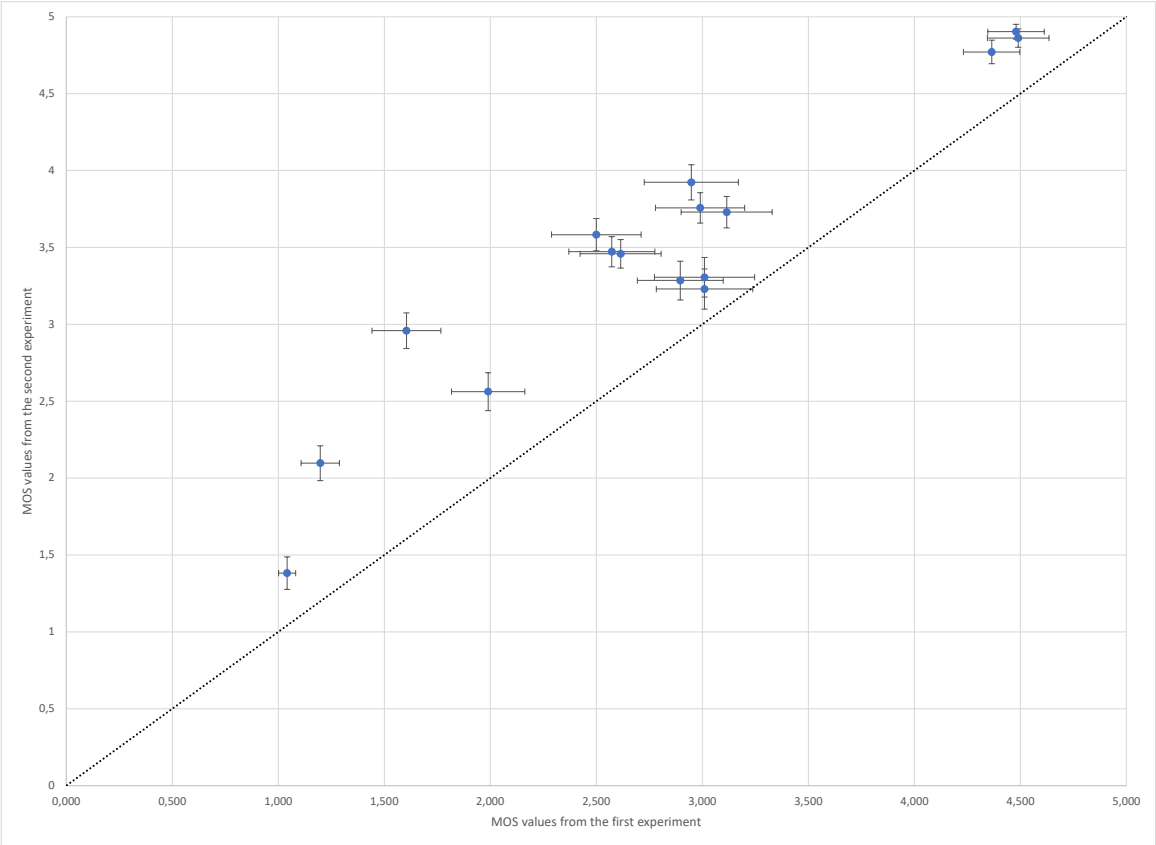


Figure 6.4: Comparison of the MOS values from the experiments.

All MOS values from both experiments are presented in Table 6.6. The differences between the MOS values are quite big. Some conditions have the difference in MOS values even bigger than 1. The conditions were rated on 5 point scale so 1 point is a huge difference. The most significant change in the MOS value between the tests happened in condition C11 which is a studio recording with added watermark with strength 200. The subjects in the first experiment rated it mostly as poor or bad with the final MOS score of 1.60. The listeners in my experiment rated it mostly as fair sample with the final MOS value of 2.96.

Condition	MOS (first test)	MOS (second test)
C01 - reference condition	4.49	4.86
C02 - watermark strength 10	4.48	4.90
C03 - Acoustic recording	3.12	3.73
C04 - watermark strength 30	4.37	4.77
C05 - Pub noise	3.01	3.23
C06 - Pub noise, watermark strength 10	3.01	3.31
C07 - watermark strength 75	2.95	3.92
C08 - HMMWV noise, watermark strength 10	2.50	3.58
C09 - HMMWV noise	2.57	3.47
C10 - Pub noise, watermark strength 30	2.90	3.28
C11 - watermark strength 200	1.60	2.96
C12 - HMMWV noise, watermark strength 30	2.62	3.46
C13 - Pub noise, acoustic, watermark strength 30	1.99	2.56
C14 - watermark strength 650	1.04	1.38
C15 - Acoustic recording, watermark strength 500	1.20	2.10
C16 - Acoustic recording, watermark strength 100	2.99	3.76

Table 6.6: MOS values from the first and second experiments.

The question is why there is such a difference in the results from the experiments. This is a question that cannot be easily answered and we can only guess. Both experiments were conducted in the same testing laboratory and all the technical specifications were the same in both tests, so different laboratory conditions like for example different headphones could not cause the difference. That means that the difference must have been caused by the subjects participating in the experiments. Participants of both experiments were evenly divided based on their gender and age so that also should not be the cause of different results.

The only difference between the subjects of each experiment was their nationality and level of their English. The participants in my experiment most likely had worse knowledge of English than the listeners in the first experiment. The subjects in my test rated the samples with higher scores. If there is a connection between these two facts is not clear.

The difference in results might have appeared because the Chinese listeners were more concentrated on understanding the meaning of the sentences spoken in the samples and therefore did not pay enough attention to the deterioration of the quality. Another explanation might be a different perception of the MOS scale. The words excellent, good, fair, poor and bad might be understood differently by a native speaker and a person who is not that familiar with English.

The difference might have also been caused by a different mentality of the participants of each experiment. It is possible that participants in my experiment just did not have such high expectations of the quality of the recordings. Their requirements of the quality were not as high as of the subjects in the first experiment so they rated them with higher scores.

However, even though the MOS values in my experiment were all statistically significantly higher than in the first test, the results concerning the level when the watermark statistically significantly decreases the quality of the recording were the same.

Chapter 7

Conclusion

Before conducting the experiment, I have studied the existing international recommendations concerning subjective testing of transmission quality. Based on these recommendations, I prepared and conducted the subjective testing experiment.

For my experiment, I followed the ITU-T P.800 recommendations [4]. After a consultation with my supervisor, we decided to use the Absolute Category Rating (ACR) method instead of the Degradation Category Rating (DCR) method. We decided so because all of the objective testing methods are only compatible with the ACR method.

In my experiment I examined the impact of the watermark embedded into different sample recordings using the subjective testing. I analysed the results and compared them with the results of the previous experiment [5]. I discovered the border values of the watermarking strength where the watermark still does not decrease the perceptual quality of the recording. Yet, it is still robust enough to fulfil its function.

Chapter 8

Future Plans

Digital watermarking is an important tool to protect digital content. And in my opinion its importance will only grow.

To further develop my experiment I could perform an objective test to see whether I can receive similar results to the subjective test. One of the objective testing algorithms is POLQA (Perceptual Objective Listening Quality Assessment) ITU-T P.863 [11].

It would also be useful to conduct another subjective testing experiment with different set of samples. During one experiment it is not possible to cover sufficient amount of testing conditions to get the exact border values of the watermarking strength. In the next experiment I would mainly focus on the values of the watermarking strengths ranking from 30 to 100 because that is where the border appears to be. For this experiment, the DCR method would be more suitable than the ACR method as it can better distinguish subtle changes in the quality.

I could also conduct a subjective testing experiment with a parallel task that can simulate a real environment. The subjects taking part in the subjective testing experiment with parallel task have to also focus on different activity apart from listening to the samples. Unfortunately this could not have been performed due to the covid restrictions.

Bibliography

[1] Yugendra Chincholkar and Sanjay Ganorkar. Audio watermarking algorithm implementation using patchwork technique. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–5, 2019.

[2] Michael H. Herzog, Gregory Francis, and Aaron Clarke. *Variations on the t-Test*, pages 51–59. Springer International Publishing, Cham, 2019.

[3] Konrad Hofbauer. *Speech watermarking and air traffic control*. Faculty of Electrical and Information Engineering Graz University of Technology, Austria, 2009.

[4] Rec. ITU-T. P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 22, 1996.

[5] Y Kowalczuk and J Holub. Evaluation of digital watermarking on subjective speech quality. 2021.

[6] Zhenghui Liu, Yuankun Huang, and Jiwu Huang. Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks. *IEEE Transactions on Information Forensics and Security*, 14(5):1171–1180, 2019.

[7] Mohammad Ali Nematollahi and S. A. R. Al-Haddad. An overview of digital speech watermarking. *International Journal of Speech Technology*, 16(4):471–488, May 2013.

[8] Mohammad Ali Nematollahi, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales. *Preliminary on Watermarking Technology*, pages 1–14. Springer Singapore, Singapore, 2017.

[9] C.I. Podilchuk and E.J. Delp. Digital watermarking: algorithms and applications. *IEEE Signal Processing Magazine*, 18(4):33–46, 2001.

[10] Victor R. Preedy and Ronald R. Watson, editors. *T-Test*, pages 4334–4334. Springer New York, New York, NY, 2010.

[11] ITU-T Rec. P.863: Perceptual objective listening quality assessment. *International Telecommunication Union, Geneva*, 2011.

[12] Martin Steinebach. *Digitale Wasserzeichen fuer Audiodaten.* Shaker, 2004.