ON THE CORRELATION OF PSYCHOACOUSTIC METRICS WITH MUSHRA-BASED SUBJECTIVE SOUND QUALITY PERCEPTION FOR TELEVISIONS

A Thesis

by

Çağlar İşlek

Submitted to the
Graduate School of Sciences and Engineering
In Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Electrical and Electronics Engineering

Özyeğin University
June 2019

ON THE CORRELATION OF PSYCHOACOUSTIC METRICS WITH MUSHRA-BASED SUBJECTIVE SOUND QUALITY PERCEPTION FOR TELEVISIONS

Approved by:

Associate Professor Cenk Demiroğlu,
Advisor
Department of Electrical and Electronics Engineering
Özyeğin University

Professor Feridun Öziş
Department of Music Sciences
Dokuz Eylül University

Professor Taylan Akdoğan
Department of Electrical and Electronics Engineering
Özyeğin University

Date Approved: 9 August 2019

To my wife and lovely daughter

ABSTRACT

Even though the loudspeaker technologies used in televisions (TVs) and objective sound quality assessment techniques had continuously advanced over the years, the literature on automatic sound quality assessment in the TV domain is still insufficient and scarce. Most of the TV manufacturers use acoustical and electrical measurement methods to evaluate the sound quality of TVs. However, in real life, how well these methods correlate with human perception has not been documented in the TV context. Quantifying those correlations may not only allow standardization between TV manufacturers but also can eliminate the cost of expensive and time-consuming subjective jury-testing processes. The aim of this study is two-folds. One goal is to propose subjective and objective test guidelines that can be used for TV manufacturers for sound quality assessment. Whereas, the second goal is to analyze the correlation between some of the commonly-used objective measures and the subjective perception of TV sound quality as measured with the MUSHRA test (ITU-1534). To that end, in this thesis we describe our carefully designed subjective and objective test setup and present the correlation analysis results.

ÖZETÇE

Televizyonlarda(TV) kullanılan hoparlör teknolojileri ve nesnel ses kalitesi değerlendirme methodları yıllar boyunca sürekli olarak ilerlemiş olsa da, TV alanındaki ses kalitesi değerlendirmesi ile ilgili literatür mevcut değildir. TV üreticilerinin çoğu, TV'lerin ses kalitesini değerlendirmek için akustik ve elektriksel ölçüm yöntemleri kullanır. Ancak, gerçek hayatta bu yöntemlerin insan algısı ile ne kadar ilişkili olduğu TV bağlamında belgelenmemiştir. Bu korelasyonları ölçmek yalnızca TV üreticileri için standardizasyona izin vermekle kalmayabilir, aynı zamanda pahalı ve zaman alan öznel jüri test süreçlerinin maliyetini de ortadan kaldırabilir. Bu çalışmada iki konu amaçlanmıştır, birincisi; TV üreticileri için ses kalitesi değerlendirmelerinde kullanılabilecek öznel ve nesnel test methodu önermektir. İkinci amaç, yaygın olarak kullanılan bazı objektif metrikler ile MUSHRA testiyle (ITU-1534) ölçülen öznel ses kalitesi algısı arasındaki korelasyonu incelemektir. Bu amaçla, özenle tasarlanmış öznel ve nesnel test yapısı tanımlanacak ve korelasyon analiz sonuçları paylaşılacak.

ACKNOWLEDGEMENTS

I would like to express my very great and special appreciation to my supervisor Assoc. Prof. Cenk Demiroğlu for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. I would also like to express my deep gratitude to Professor Feridun Öziş for his advice and assistance in subjective test sessions.

My grateful thanks are also extended to my manager Mr Eyüp Karaoğlu for his encouraging and supportive approach, to Mr Özkan Ayyıldız, who helped me to measure electrical performance of Televisions, and to Ms Şebnem Ecem Koçak for her support to review my thesis.

I would also like to extend my thanks to the academic members of the Dokuz Eylül University Department of Music Science for participating in long-term jury tests, to the acoustic laboratory members of Vestel Elektronik A.Ş for allowing us to use all the facilities they have, to Scientific and Technological Research Council of Turkey (TÜBİTAK) for supporting this thesis(Project No: 3160398).

Special thanks to my mother, father and sister for supporting me in every stage of my life. Without their support, I would not have been able to achieve my goals. Little thanks to my lovely daughter at least for letting me study on the thesis at nights.

Finally, I would like to offer my heartily thanks to my wife for all her sacrifice and patience during my thesis. Thanks to her to being my wife, to stands with me on my good and bad days and sharing my troubles and happiness without any regrets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In this chapter an overview, the background, aims and the objectives of the study are presented.

1.1 Introduction

When sound reproduction systems were first released decades ago, the purpose of the design was to keep noise and harmonics level produced by the product at a minimum level. The assumption was that the lower the noise the better. Further studies bred the importance of the frequency effect on the sound. The assumption was that the flatter the impulse response the better. After those investigations, the taste of the sound notion appeared and there was a need to not only look at the noise level, frequency response or other basic terms but also to look at the detailed quality of the sound. Later, it was accepted that the term sound quality had both subjective and objective side since it could be described by subjective representations or it could be measured with objective methods, yet none of them was enough if applied separately. Zwicker came up with the idea of loudness based sound quality metrics, which defines the subjective perception of sound pressure[2]. Since then, many objective measures based sound quality studies appeared which used loudness based sound quality metrics to define the accuracy, fidelity, or intelligibility of the sound, however, we believe that there is still untouched fields on the research topic "Sound Quality".

1.2 Background of the Study

Because speech and audio are produced by many commercial products electronically, measuring the quality of sound has been an important challenge for decades. Subjective jury testing is commonly used for assessment of audio quality and has been standardized by International Telecommunication Union (ITU) in 1994. An ITU standard for assessment introduces a method for comparison of high-quality audio systems which has very small differences [3]. Another popular standard published in 2001 by ITU [4] is called "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)" and is intended to compare multiple products with a reference signal. MUSHRA method is preferred when confidence intervals of the products are very low and it returns significant results if the number of test subjects is reasonable. Even though subjective assessments are standardized and give reliable and repeatable results, it is expensive and time-consuming to perform those tests on each product. Therefore, investigation of objective psycho-acoustic metrics and correlation studies between objective metrics and subjective assessments have been an active research area [5, 6, 7].

Among the popular objective metrics, Zwicker's loudness model [2] is thought to be a milestone on sound quality studies. Since the goal of psychoacoustics is to understand how people perceive and experience the sound, there is also a large literature about binaural perception studies [8, 9, 10, 11, 12].

1.3 Statement of the Research Problem

In recent years, as a result of the growth of content providers, video compression algorithms and developments in display technologies make customers expect "home theatre" experiences. The developments on the display side also triggered more investigation and improvements in the audio quality that further increased the audio quality expectations of the consumers. Moreover, as TV screens and cabinet sizes got

thinner, available space for loudspeakers became smaller, which made high-quality audio production harder. Not only loudspeakers but also other layers that affect the sound performance exist; namely, the main integrated circuit (IC) for decoding and post-processing, power amplifier and the TV cabinet. All those factors cause linear and nonlinear distortions that affect the perceptual quality of music and speech signals rendered on TV [13, 14].

Even though TV manufacturers apply many tests [15, 16, 17, 18] on their products for certification and verification purposes, there is no widely-accepted standard for objective assessment of perceptual sound quality in the TV industry. That, in turn, leads to the utilization of expensive in-house listening tests or certification through centres that are specialized in subjective assessment of sound quality. Because of the issues discussed above in the TV context, more work is needed to assess how well existing objective measures correlate with the listening tests in the context of the speech, music, and noise sounds. Moreover, there could be many different versions of the product in the design phase, so these tests have to be repeated many times even during the design phase. The motivation of this study comes from the need for automatic prediction of the subjective preference of the customers on the audio quality of TVs.

1.4 Aims and Objectives of the Research

The aim of this study is two-folds. One goal is to propose subjective and objective test guidelines that can be used for TV manufacturers for sound quality assessment. The second goal is to analyze the correlation between some of the commonly-used objective measures and the subjective perception of TV sound quality as measured with the MUSHRA test. These two folds will enable better prediction of the sound quality on TV products. Figure 1 describes the process to determine the sound quality of TV products before/after this study. Once correlation analysis is done, the results

can be used to predict the sound quality of the TV products.

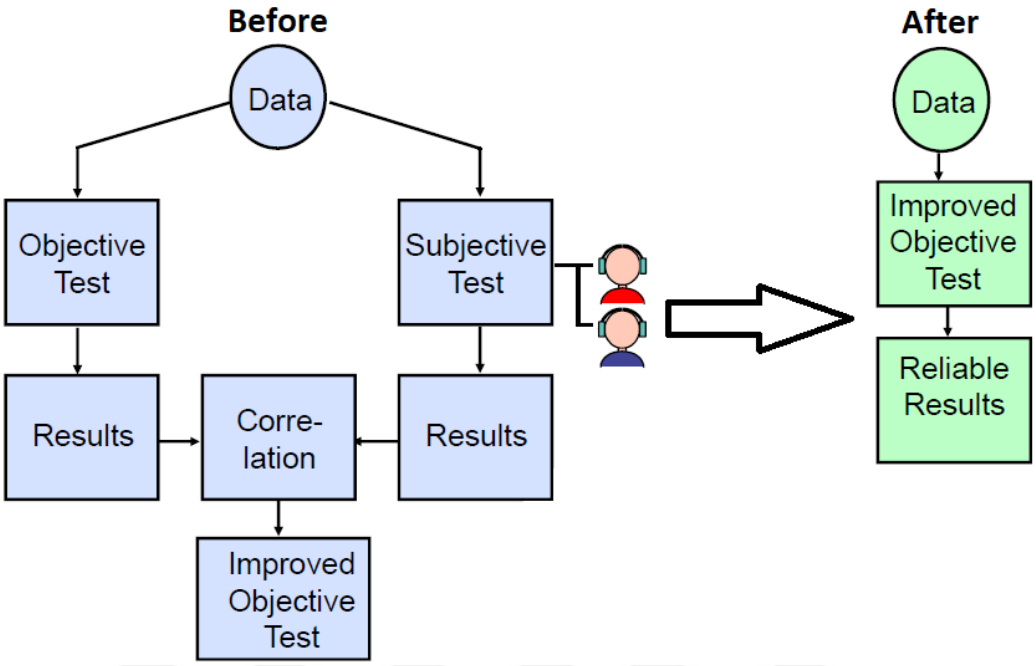

Figure 1: Flow chart of the measurement process before/after this study

To that end, we describe our carefully designed subjective and objective test setup and present the correlation analysis results. We first describe the setup that we designed for performing the MUSHRA test on 6 different TVs with different amplifier-loudspeaker pairs. Then, we present the results obtained with some of the more commonly used objective measures and an analysis of the correlation between the objective and subjective test results. Predicting an objective measure of the sound quality of TV is done in below consecutive phases;

- Experimental design: in order to obtain reliable results this phase has substantial importance. Determining the test TV (t-TV) samples, selecting listeners, the definition of test signals (music, movie, speech, etc.) and other experimental factors will be discussed in this part.

- Performing MUSHRA test: Challenges of performing MUSHRA test on TV products will be discussed and an HTML based open-source test application

[19] is modified for this study will be presented.

- Compute psychoacoustic metrics: 11 different psychoacoustic metrics computed for all recordings captured from t-TVs.

- Correlation Analysis: The Pearson correlation coefficient was used in the analysis of the correlation between subjective scores and objective scores for different TVs.

1.5 Research Questions

In this study we are looking for answers to below questions;

- Which subjective metrics should be used to accurately identify the preferences and complaints of customers?

- Are the current objective metrics suitable to study sound quality on TV products?

- How to define subjective and objective test guidelines that can be used for TV manufacturers for the sound quality assessment?

- What are the correlations between pre-defined objective and subjective metrics.

CHAPTER II

LITERATURE REVIEW

An overview of the literature is presented here. Besides, some most relevant studies in this field are presented in detail.

2.1 Measurement Basics on Sound Sources

Sound quality is a performance criterion for all devices that produce sound. In this context, sound quality studies for many different devices are available in the literature. When the studies in this area are reviewed, it is possible to find at least several studies about sound quality analysis for almost any electronic or mechanical device that is a sound source. Sound quality studies can be divided into two different categories: devices that produce meaningful sounds, and devices that produce noise.

2.1.1 Studies on noise generating devices

The aim of the sound quality studies in this field is to reduce the produced noise level, to measure the noise accurately, and to determine which forms of noise are less disturbing to people. The answer to the question of how to make the noise less disturbing to people has been the outcome of the studies in this field.

The car industry is one of the areas where sound quality studies are carried out meticulously. Although the sound quality studies on car engine sound have been continuing for years, there are still new publications on this subject[20]. When the studies in that area are examined, it is seen that there are studies even on the sound of the car door closing[21]. One of the other areas where sound quality studies matter is consumer electronics products. These products are mostly used in our daily lives and have sound sources. Sound quality studies have been examined on the washing

machines[22], the hairdryer[23], the vacuum cleaner[24] and many other products.

2.1.2 Studies on devices that generate meaningful sounds

Most of the devices in this category contain a sound reproduction system. Basic elements in this system are: an audio Digital Signal Processor (DSP), a power amplifier and one or more loudspeakers. This system, called sound reproduction, takes an audio source file (mp3, mp4, wave etc) as input and generates sound waves as output. The input audio source file is converted to an analog signal after passing through the DSP, then the level of the analog signal is increased on the power amplifier and transferred to the loudspeakers. Finally, the amplified analog audio signal is converted into a sound wave on the loudspeaker and the process is completed. Televisions, multi-channel sound systems, telephones are examples of devices in this category.

The most important factors affecting the sound reproduction system is bit rate and sample rate of the audio source file, audio amplifier's electrical performance and loudspeaker's acoustical performance. Bit rate and sample rate parameters are external parameters that affect the sound reproduction system. Research shows that the sample rate and bit rate parameters have a significant effect on sound quality[25, 26]. It is recommended that the audio source file with good sound quality has a bit rate of at least 64-128 kbps, the recommended value is 96 kbps and above. Since the human hearing bandwidth is 20Hz-20kHz, an audio sample rate above the 40kHz (Usually 44.1KHz) is preferred. All audio source files used in this thesis were recorded as 192 kbps bit rate and 44.1 kHz sample rate. Since this project aimed to investigate the sound quality of sound reproduction systems (TV-specific), the effect of bit rate and sample rate parameters on the sound quality were ignored.

There are studies and even standards in the literature for the measurement of an audio amplifier and loudspeaker performances which are the two most important

parameters of the sound reproduction system. Audio power amplifier electrical performance tests are given in the CEA-490-A R-2008 [27] standard with details. All of these tests are electrical tests and the most commonly used ones are as follows;

- Power output rating in Watts
- Dynamic headroom in dBr
- Total harmonic distortion plus noise (THD + N) in percent
- Signal to noise ratio (SNR) in dBr
- Deep Noise in dBr

In this thesis, two different audio amplifiers were used, both of which have specification values in electrical tests. The purpose of using two different amplifiers was to observe the effect of amplifier performance parameters on sound quality.

The loudspeakers, the last component of the sound reproduction system, are the most important part of the system in terms of sound quality. In this thesis, four different loudspeaker sets were used. For acoustic performance tests on loudspeakers, IEC 60268-5 [28] is used as a reference document where the flatness of the acoustic frequency response is the key performance parameter. The main tests can be listed as follows;

- Flat frequency response (+/- 3 dB) between 250Hz-8kHz
- Effective Frequency range (+/- 10 dB) between 100Hz-18kHz
- Sensitivity in dBrA
- Signal to noise ratio (SNR) in dBr
- Acoustical THD + N

All of the above electrical and acoustical tests were performed for the televisions that were used in this thesis and were the key criteria for the selection of televisions. You can find detailed results and analysis for each Television in section 3.2.

Studies and standards given above can be considered as a prerequisite for sound reproduction systems. It is not possible to talk about the concept of sound quality for products that do not achieve the performance criteria described in these standards. These standards and tests are essential but not sufficient to analyze products sound quality. The concept of sound quality is still a challenge that one encounters after providing the aforementioned specifications. Our preliminary studies have shown that there are differences in sound quality between different televisions that meet the defined specifications in the standards. In order to measure these differences and present them with objective methods, it is seen that in-depth and specific test methods are needed. In this thesis, especially the experimental studies on televisions will be carried out to measure the sound quality differences by going beyond traditional methods.

2.2 Related Sound Quality Studies

Since the objective test methods mentioned in the previous section are not sufficient to distinguish the sound quality performance of TV products, recently studied objective methods will be present in this section.

Perceptual Evaluation of Speech Quality aka PESQ[11] known as base standard (ITU-T recommendation P.862) to predict the speech quality perceived by human beings. Even though it is mostly used to predict speech quality of the telephony systems it can be also applied to any speech recording signal. The PESQ method receives the recorded speech signal as an input and compares it with the original signal, it outputs a prediction of mean opinion score(MOS) based on a scale from 1(bad) to 5(excellent). Perceptual Objective Listening Quality Analysis(POLQA)

[12], known as successor of the PESQ, presents an advanced and extended model of PESQ. POLQA provides an option to analyze higher bandwidth speech signals. Although the test methods such as PESQ and POLQA work are effective, they are not sufficient for the tests of TV products by themselves. These methods are used only for the analysis of speech signals, whereas TV is a system that can produce many different sounds such as music, film, speech etc.

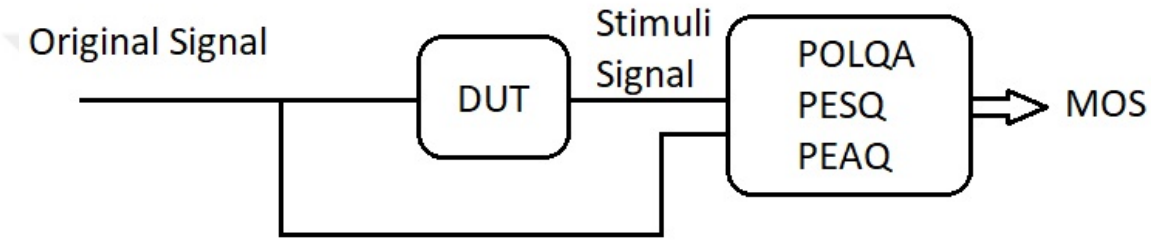

Figure 2: The basic principle of PESQ, PEAQ and POLQA algorithms

Another standardized method, Perceptual Evaluation of Audio Quality (PEAQ)[10], can be considered as the most relevant method. The method proposed by PEAQ is to estimate the sound quality based on the comparison of the input and the reference signal as shown in Figure 2. The major difference of PEAQ with PESQ and POLQA is that it can be applied to all audio signals, and in this respect, it is similar to our study. When the PEAQ method is examined in all details, it is seen that there are 2 different deficiencies. First; Since the PEAQ algorithm is based on a dual subjective comparison method (ITU 1116-1), it evaluates the performance of each product relative to the reference signal. If it is desired to evaluate the performance of multiple products relative to each other, it may not give reliable results. Secondly, methods such as PESQ, POLQA and PEAQ give a score on a 5-point scale that defines the sound quality of the product as an output. With these algorithms, it is possible to obtain information about how satisfactory and good the sound is to be tested, but it is not possible to make a detailed comparison. For example, in the case where

the signal under test is worse than the reference signal, it is not possible to obtain information on what makes it worse. Thus, alternative methods are still needed if the tested product is to be examined in the technical layer like bass - treble balance, loudness, speech quality etc.

None of the aforementioned methods provide a cheap, reliable and stable solution to the sound quality measurement problems of TV manufacturers. According to the information received from the after-sales service departments of many TV brands, complaints on sound quality for TVs can be categorized under 4 different headings; bass balance, treble balance, speech quality, overall sound quality. For this reason, most of the TV sound settings allow for compensating small effects on those parameters. The method proposed by this thesis has provided a solution to the needs of TV manufacturers by estimating the above four main subjective metrics over objective metrics. To that end, we used MUSHRA methodology and loudness based psychoacoustic metrics since MUSHRA provides flexibility to compare multiple products and allows comparing small differences.

CHAPTER III

EXPERIMENTAL SETUP

ITU standards [3, 4] are used as guidance for designing the experimental setup and subjective tests. There are several challenges in performing listening tests with TV sets. In this section, technical difficulties and limitations will be discussed along with the description of our test setup.

3.1 Design of the Test Signals

The TV is a media playback device on which one can listen to any type of content. Still, for the most part, the content can be broadly categorized as music, movie and speech signals, similar to EBU test on audio codecs [29]. Thus, our audio samples are chosen from those domains. In MUSHRA tests, the duration of each test signal is required to be below 10 seconds and each grading session should be less than 30 minutes. To fulfil those requirements 17 audio samples are used, each of which is approximately 10 seconds. Six of those samples are clean speech samples without any other tone or effects, 3 of them is male and other 3 is female, all speech signals defined in the Turkish language since the native language of the listeners are Turkish. Ten of the samples are music samples including rock, pop, classic and some regional kinds of music. Four of them are sound effects from movies without speech. Table 1 gives details about the audio sample structure. With given audio sample structure the average total duration of the subjective test was close to 40 minutes, during the tests, to prevent distress and fatigue on listeners subjective test was split. In order to eliminate sample rate and bit depth effects on the assessments all test signals and recordings from TVs set to the bit depth of 16 bits/sample and sampling frequency of 44.1kHz.

Table 1: Description of test contents for each subjective test

Subjective Metrics	Number of Samples	Genre of Samples	Duration of Samples
Bass Balance	3	Pop, Rock Hip-Hop	60 sec
Treble Balance	3	Jazz, Pop Electro music	60 sec
Speech Quality	6	3 female voice 3 male voice	98 sec
Overall SQ	5	Movie scene, Jazz Pop, Classic, Country	99 sec

3.2 Configuration of t-TVs

Products with different sound qualities were selected as t-TVs. To focus only on the amplifier and the loudspeaker, all software-based post-processing features on TVs were disabled excluding default parametric equalizer settings defined by acoustical engineers to adjust flat frequency response for each t-TV [30]. Table 2 shows the combination of amplifiers and loudspeakers configuration for each t-TV sample. Since all TVs belong to the same brand, features such as main IC, software version, panel size, mechanical effects, etc. that may affect the sound performance are identical for all TVs. Thus the only standing differences between t-TVs are loudspeakers and audio amplifiers.

List of t-TV amplifiers was categorized as either low-cost [31] or high-performance [32]. Since most of the audio amplifiers have sufficient performance on electrical testing we limit the amplifier combinations to two. As seen in Table 2 only t-TV 1 and t-TV 4 has a high-quality amplifier model. That will allow us to compare amplifier effect on subjective results. Subjective comparison between t-TV 1 versus t-TV 3 and t-TV 4 versus t-TV 6 provides the data if is there any noticeable difference between amplifier models.

Table 2: t-TV's audio amplifier and loudspeaker configuration

Sample No.	Audio Amplifier	Loudspeaker
t-TV 1	High performance	Loudspeaker A
t-TV 2	Low cost	Loudspeaker B
t-TV 3	Low cost	Loudspeaker A
t-TV 4	High performance	Loudspeaker C
t-TV 5	Low cost	Loudspeaker D
t-TV 6	Low cost	Loudspeaker C

Table 3: Specifications for Loudspeaker A,B,C and D. Since model D use same loudspeaker with model C only subwoofer part is described model D

	A	B	C	D
Rated Impedance	8 +/- 1.2 ohms	8 +/- 1.2 ohms	8 +/- 1.2 ohms	8 +/- 1.6 ohms
Power Rating	10 watts	12 watts	12.5 watts	12 watts
Resonance Frequency	240 +/- 48 Hz	240 +/- 48 Hz	200 +/- 40 Hz	150 +/- 30 Hz
Dimensions	29/33/140 mm	29/33/140 mm	25/50/141 mm	9/160/240 mm
SPL	80 +/- 3 dB	82 +/- 3 dB	83 +/- 3 dB	79 +/- 3 dB

Specification values for four different types of loudspeaker sets are shown in Table 3. Loudspeaker A is a very low-cost loudspeaker model, Loudspeaker B is also a low-cost model with a larger magnet than Loudspeaker A, Loudspeaker C is mid-range speaker enclosed in a box, and Loudspeaker D has the same loudspeaker set that is in Loudspeaker C with an additional subwoofer. Dimensions of the loudspeaker units are also given in Table 3, as can be seen in the table thickness of the loudspeaker units is varying between 9 to 29 mm. Loudspeaker dimensions affect the product sound quality, bigger the loudspeaker size provides more gain in the low-frequency band. Since most of the TV products have flat LCD panel or OLED and QLED panels recently they have extremely thin back cabinets. Thus there is no enough volume for

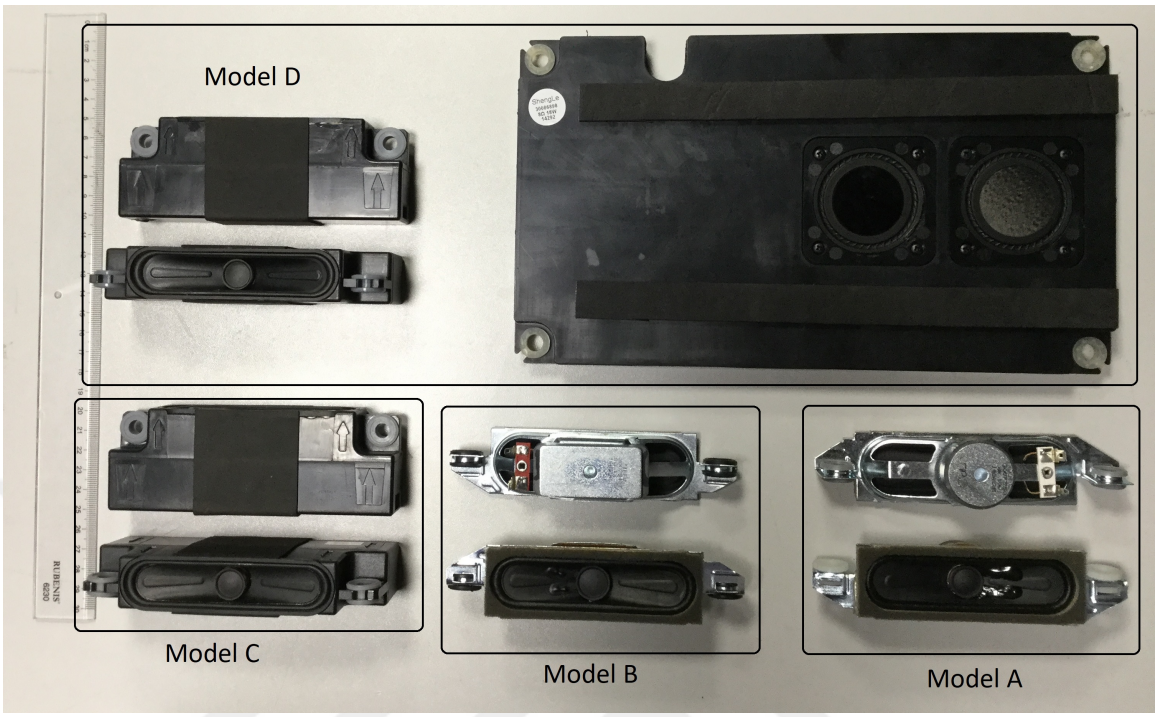

Figure 3: Loudspeaker units used with t-TV samples.

loudspeaker units. This issue is known as a root cause of the sound quality problem on TV products. t-TV samples prepared for this thesis has also a thin back cover issue since all of them has an LCD panel. Loudspeaker units used in t-TV samples are shown in Figure 3.

3.2.1 Electrical test results of t-TVs

Electrical performance tests were performed before starting sound quality studies with all TVs selected for testing. All measurement are captured with an advanced audio analyzer test device[33]. These tests are performed to measure audio amplifier performance. It is not meant to start acoustic tests and sound quality tests with a product whose electrical test results are not at the desired level due to any issues in hardware design. As a prerequisite, the electrical test results must comply with the specification values given in Table 3. Detailed results are also shown in Table 3. It is observed that all t-TVs achieve the desired performance scores. When the results

of t-TV 1 and t-TV 4 are considered, it has a better performance than other t-TVs. This difference comes from audio amplifier preference on t-TV since t-TV 1 and t-TV 4 has better quality audio amplifiers it is expected to have better results on these t-TVs. In order to eliminate any loudness differences all t-TVs adjusted to output power at the level of 10 Watts. Dynamic range describes the ratio between the lowest and highest signal level that the amplifier can provide.

Table 4: Electrical test results of t-TVs

Sample No. Unit Specification	Power Output Watt 10 +/- 1	Dynamic Range dBr > 70	THD+N % < 2	SNR dBr < -50
t-TV 1	10.5	82	0.3	-58
t-TV 2	10.3	75	1.1	-52
t-TV 3	10.6	75	1.2	-53
t-TV 4	10.2	83	0.2	-59
t-TV 5	10.1	76	0.9	-52
t-TV 6	10.0	74	1.1	-52

Although all t-TV samples provide the desired value, t-TV 1 and t-TV 4 provide more dynamic range due to amplifier difference. THD+N describes the ratio between fundamental signal and harmonics plus noises when the output power level is set to maximum. SNR and Deep Noise describes the noise density compared to the fundamental signal when the input level is low. Again t-TV 1 and t-TV 4 provide better performance on THD+N, SNR and Deep Noise. The reason we prefer to use a more powerful amplifier with 2 of t-TVs is to discover if these small differences can be detected by listeners or not.

3.2.2 Electro-acoustical test results of t-TVs

Acoustic tests were performed to measure the performance of loudspeakers used on t-TVs. All electro-acoustic test results are given in Table 5. Measurements were captured with an omni-directional microphone [34] positioned at 1 meter away from the

TV, Figure 4 shows the environment for electro-acoustical tests. The measurement setup also includes an audio interface [35] and measurement software [36]. Measurements were taken in a fully anechoic chamber. Sensitivity represents the maximum SPL that can be reproduced by loudspeaker itself. As seen in Table 5, the product with the highest sensitivity is t-TV 5, which, unlike other products, has one subwoofer, which is expected to be high. Sensitivity is a major parameter to define the power of the loudspeakers but it won't be this studies concern. Acoustical THD + N represents a ratio between fundamental signal and harmonics plus noise(room effects and loudspeaker internal noise) which should be measured at the maximum volume level. All t-TV samples have similar and sufficient result for THD+N, t-TV samples which have high-quality audio amplifier models have better results on THD+N. Our preliminary studies have shown that products with THD + N values of more than 2% disturb the listeners. Therefore, this criterion was taken into consideration when deciding which loudspeaker to use.

Table 5: Electro-acoustical test results of t-TVs

Sample No. Unit	Flat Response N/A	Effective Response N/A	Sensitivity dB	THD+N %
t-TV 1	✓	**x**	84	%0.8
t-TV 2	✓	**x**	82	%1.4
t-TV 3	✓	**x**	81	%1.6
t-TV 4	✓	**x**	89	%0.9
t-TV 5	✓	**x**	92	%1.8
t-TV 6	✓	**x**	87	%1.9

The flatness of the frequency response of the t-TV samples could be considered another base requirement which needs to be ensured by all the t-TV samples. The requirement here is to have a stable(within 3dB range) response between 250 Hz to 8 kHz, as seen in Figures 5 to 10 all t-TV samples meets the requirement for flat response. An extended version of the flat response is effective frequency response

requirement, which is mandatory to have a flat response(within 3 dB range) between 100 Hz and 18kHz. In the sound reproduction industry, this requirement is applied to high-end products such as sound systems, monitor loudspeakers etc. Since most of the TV products can be considered as mid-range sound system, it is preferred but not required to meet specifications on the effective frequency range.

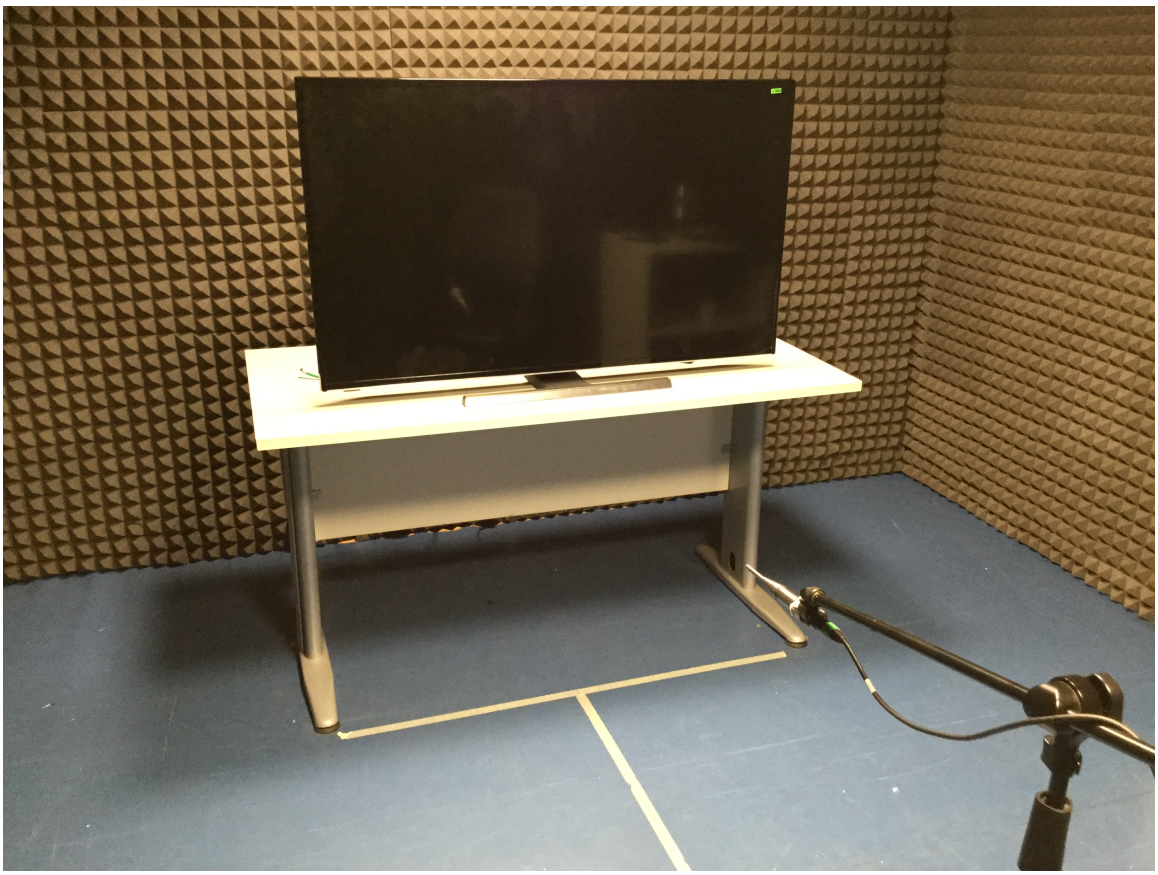

Figure 4: A sample TV from impulse response measurement session

Figures 5 to 10 show the impulse response results of products from t-TV1 to t-TV6 respectively. It is observed that t-TV 5 which has loudspeaker D has flatter frequency response than others. Especially for lower frequency bands, it has more gain due to subwoofer presence. t-TV 4 and t-TV 6 has similar curves since they have the same loudspeaker units, C. It can be also observed from the figures that t-TV 4 has more stable curve than t-TV6 due to audio amplifier effect, for some reason t-TV 1 better than t-TV 3. t-TV 3, 2 and 1 have similar frequency response curves since they have

similar loudspeaker units, A and B.

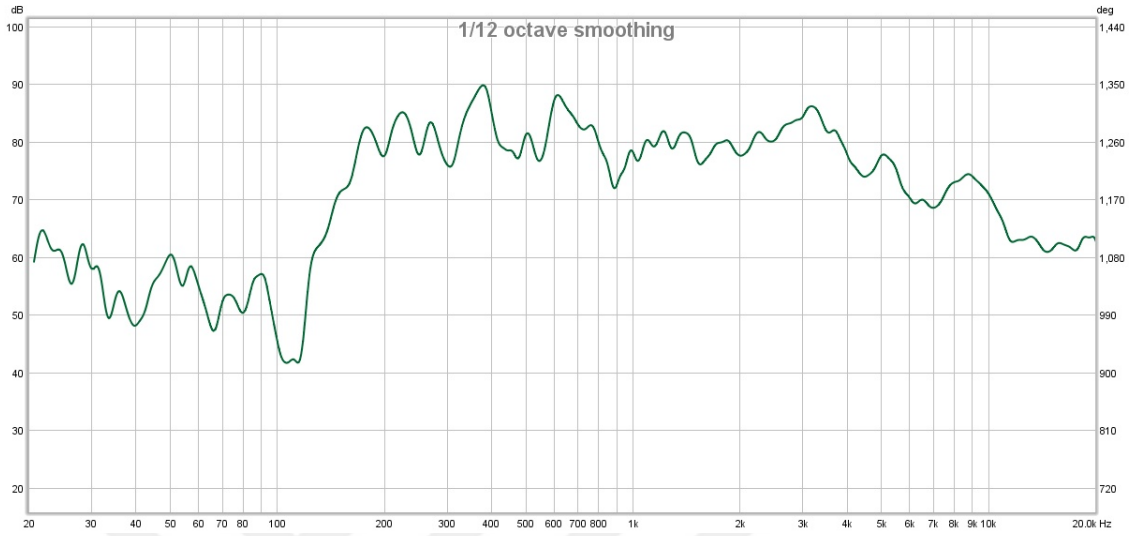

Figure 5: Impulse response line for t-TV 1. The figure includes the response between 20 Hz to 20 kHz.

Figure 6: Impulse response line for t-TV 2. The figure includes the response between 20 Hz to 20 kHz.

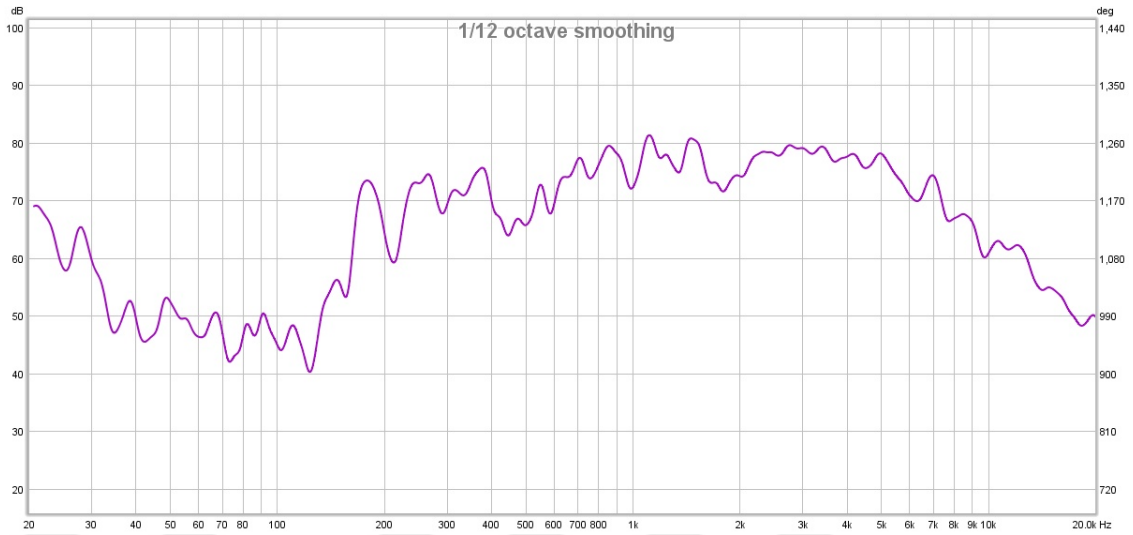

Figure 7: Impulse response line for t-TV 3. The figure includes the response between 20 Hz to 20 kHz.

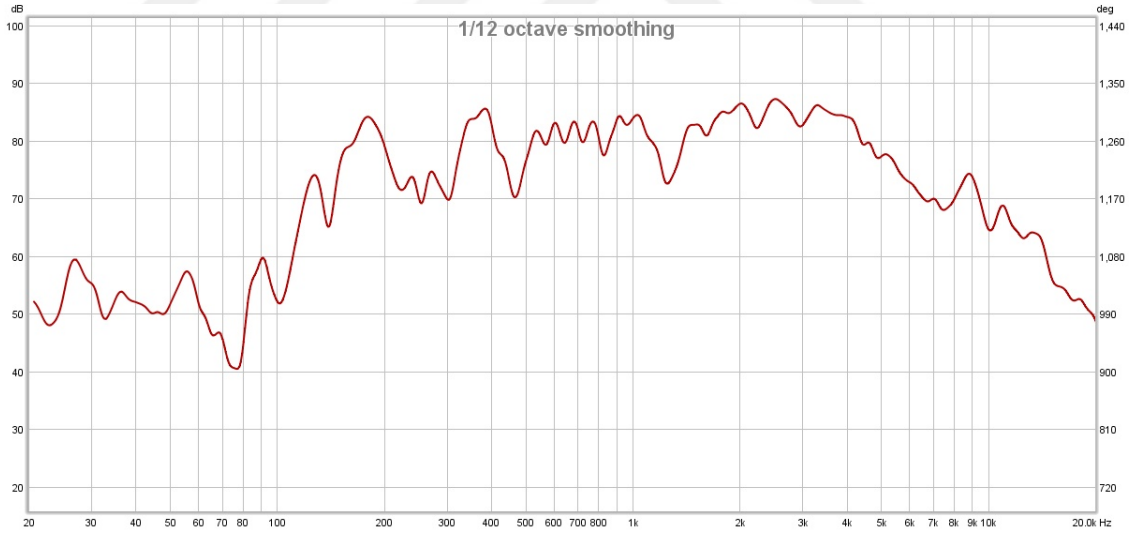

Figure 8: Impulse response line for t-TV 4. The figure includes the response between 20 Hz to 20 kHz.

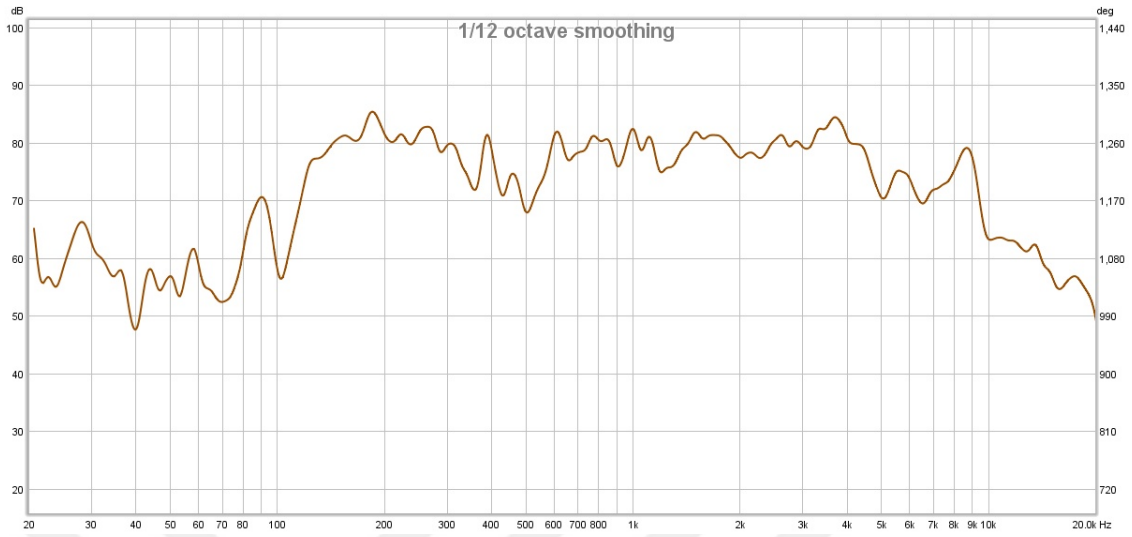

Figure 9: Impulse response line for t-TV 5. The figure includes the response between 20 Hz to 20 kHz.

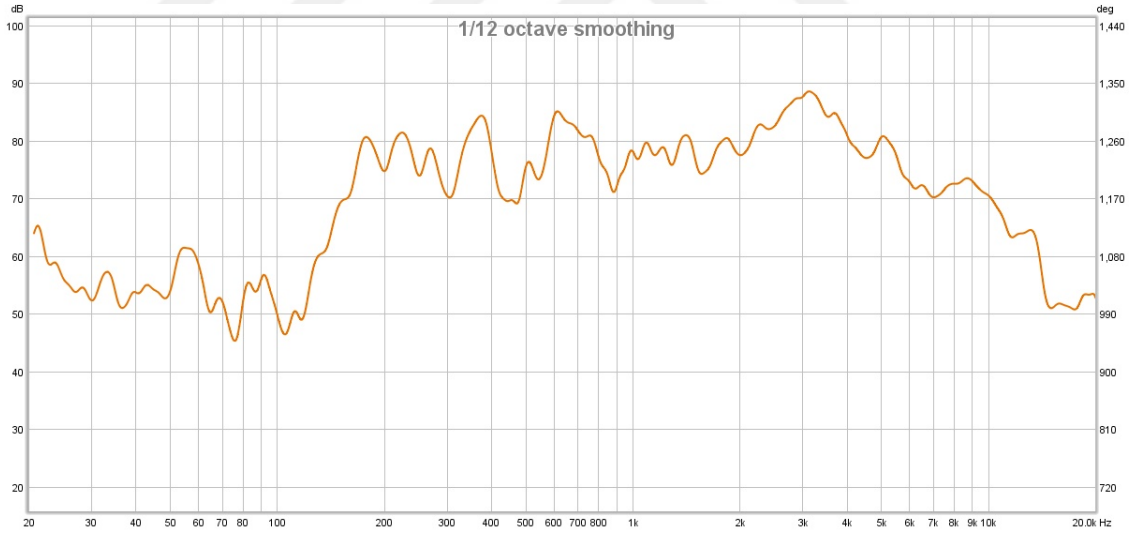

Figure 10: Impulse response line for t-TV 6. The figure includes the response between 20 Hz to 20 kHz.

3.3 Subjective Test Method

MUSHRA test is used here because differences between different TVs were expected to be small and all TVs should be assessed simultaneously. In MUSHRA, participants are presented with a reference audio sample, several test samples, a hidden version

of the reference sample, and anchor samples. Anchor sample is a severely distorted version of the reference signal. Since all test samples are presented at the same time one can obtain more significant results with fewer participants with the MUSHRA method. Also, the grading scale (0-100) of this method makes it possible to rate small differences.

In the MUSHRA test, it is recommended that participants make assessments on recordings taken from all t-TV samples (stimuli signals) for each test signal. In the case of TV, it is logistically challenging to properly place the 6 TVs in the listening room without biasing the participant or distorting the acoustic response of the room. Having enough space for TVs is also a problem. Moreover, randomly changing the t-TV order for each new stimuli is hard because the listener can most likely tell which TV is active. Therefore, we recorded the samples from each TV in an acoustic chamber and participants took the test using headphones. Details of the recording procedure are described in Section 1.5 [37].

An open-source HTML5 based framework (BeaqleJS) [19] is used as a test platform in this thesis. figure 11 shows a snapshot of the sample test screen. Each participant scored eight stimuli signals for each listening session given a reference signal on the top. Six of eight samples are stimuli signals from the t-TV samples, one of them is the hidden reference signal that is same as the reference signal presented on the top, and one of them is the hidden anchor signal which is a low-pass filtered version of the reference signal. The low-pass filter has a cut-off frequency at 3.5 kHz, +/- 0.1 dB maximum pass-band ripple, 25 dB minimum attenuation at 4 kHz. and 50dB minimum attenuation at 4.5 kHz as described in the ITU standard.

3.4 Description of Participants

20 expert listeners, ages between 19 and 46, participated in the listening tests. To get more reliable results with less number of participants, expert listeners were selected

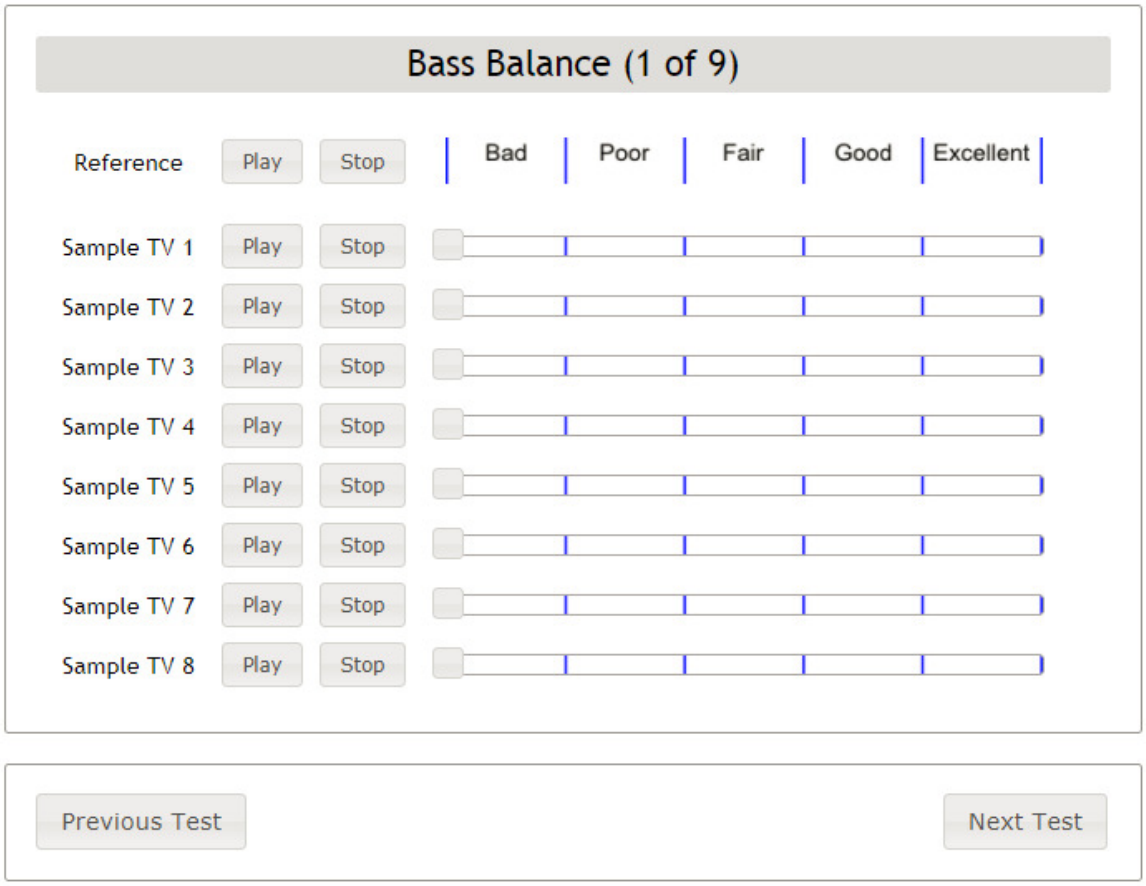

Figure 11: Html based MUSHRA test interface, sample screen of a listening session.

to take the tests in this thesis [38]. All participants are an academic member of Musicology and Music Technology department of Dokuz Eylul University. Because they all study and play music as part of their profession, they have expert-level knowledge of sound quality terminology. Since trained listeners perform more reliably than untrained listeners [39], an orientation session was performed before the actual test for each subject. Hence, all listeners were familiar with the test signals and the MUSHRA test tool before the test.

3.5 Sound Reproduction Setup

Recordings of stimuli signals were made in a listening room in accordance with the MUSHRA standard. The measurement setup described in figure 12 was used to record

each test signal from all t-TV samples. An artificial head is placed one meter away from the front side of the TV. Because head-related transfer function (HRTF) [40, 41] has a major effect on perceived sound quality, an ear type binaural microphone set (Bruel - Kjaer Type 4101-B [42]) is placed on the ear reference point of the artificial head and used as a microphone device. Test setup also included a high-performance recording device (Bruel - Kjaer Time Data Recorder Type-7708[43]) that can obtain binaural recordings.

Before each recording session, the pre-calibration of t-TV samples are done. t-TV samples are tuned to reproduce 75 dBA sound pressure level(SPL) from a one-meter distance. During fine-tune process t-TV samples are driven with -12 dBFS white noise signal. Since each t-TV sample has separate volume control with remote controller, t-TV samples are set to specific volume levels to reproduce required SPL levels. 75 dBA SPL is considered as loud enough to obtain clear recordings without any clipping and distortion. At the end of the recording sessions, 17 test signals from 6 t-TV samples resulted in a total of 102 stimuli signals. Subjective assessments of the participants and the computation of psychoacoustic metrics were performed with those stimuli signals.

Stax SR-L700 [44] headphone set is used for all listening sessions and all listeners use the same headphone set for subjective assessments. The headphone and other equipment used in the listening tests were high-quality studio equipment.

figure 13 defines sound path differences between normal playback and headphone playback. Transfer functions of the recording and playback devices distort the stimuli signals and their effects should be removed. Also, the effect of HRTF should be added into account for headphone playback case. In our test setup, the effect of the electrical response of the recording system is removed by applying the inverse response of the filter to all stimuli signals. The inverse filtering is done automatically by the recording hardware. Similarly, the inverse response of the headphone was

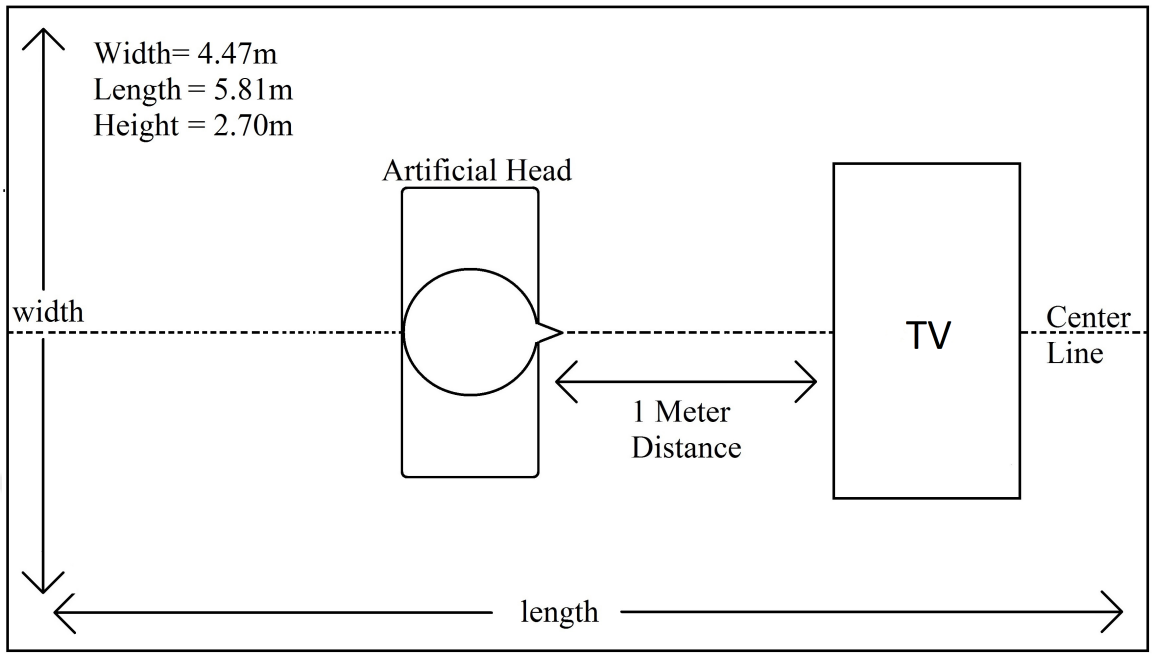

Figure 12: Binaural recording setup with an artificial head in an ITU 1116-1 Listening Room. The room has 30 dBA average background noise level and meets Noise Curve 15 specifications based on ISO Recommendation R1996 (1972). Average reverberation time is 0.25ms.

obtained from the specification sheet of the headphone manufacturer and applied to the stimuli to remove the effect of the headphone transfer function. HRTF for our recording system is also included by Bruel - Kjaer Time Data Recorder Type-7708 device.

3.6 Metric Preferences

3.6.1 Objective Measures

We used the objective measures that are available in the commercial software package PULSE by Bruel-Kjaer [45] which is capable to calculate 8 different objective metrics. In addition to these 8 metrics, we decided to calculate PESQ scores for each speech stimuli signals. Thus, in total, 9 different objective metric considered, table 6 defines those objective metrics. Our preliminary studies indicate that only 5 of them are most relevant for sound quality assessment in TVs. Detailed description for these 5

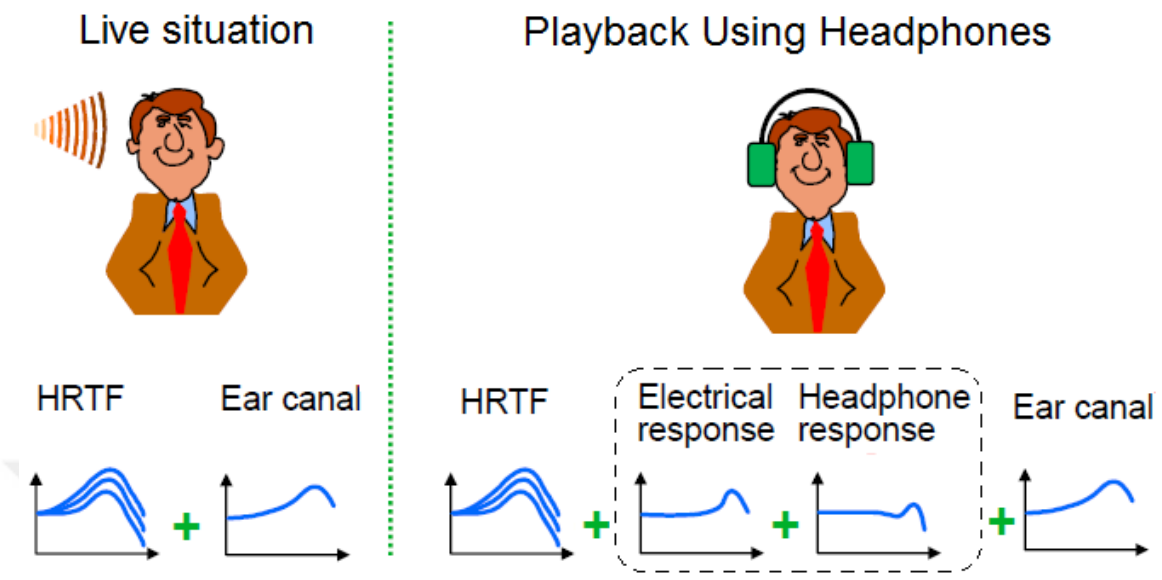

Figure 13: Sound path comparison between live situation and headphone playback cases [1]

metrics are listed below;

Table 6: Objective metric list

Metric	Unit	Standard
Fluctuation Strength	vacil	ISO 532-1:2017
Articulation Index	%	ANSI S3.5-1997
Loudness Level	phon	ISO 532-1:2017
Prominence Ratio	dBA	ANSI S1.13
Roughness	asper	ISO 532-1:2017
Sharpness	acum	ISO 532-1:2017
Tone Level	dBA Pa	ANSI S1.13
Tone to Noise Ratio	dBA	ANSI S1.13
PESQ	N/A	ITU-R P.862 (02/01)

- Loudness Level: Loudness is known as a useful metric to describe the overall sound quality of the industrial products [18]. The ISO 532-1 for (Zwicker's model) [2] is used here to measure loudness. The phenomena behind the loudness level are to compensate the all frequency content in the sound and present

a level which loud equally as 1 kHz tone. For example a sound with loudness level 70 phon means as loud as 70dB in 1 kHz tone.

- Sharpness: Sharpness of the sound is a measure of the ratio of the high-frequency content of a sound to its low-frequency content. Sound becomes brighter and when the high-frequency content, hence sharpness, increases. There is no standard for measuring sharpness. We used the algorithm described in [2] for measuring it.

- Roughness and Fluctuation Strength: Sound reproduction systems typically introduce amplitude modulation on the signal that may affect the perceived sound quality. Moreover, it is important that the system can reproduce the existing natural modulation on sounds. Since human speech organ generates sound with dominant fluctuations relationship between this metric and naturalness of speech signal are related. Roughness measures the perceptual effect of rapid (15-300 Hz) amplitude modulations. Unit of roughness is asper, which is defined as the roughness of a 60 dB 1kHz tone amplitude modulated at 70Hz [2]. A similar measure, fluctuation strength, quantify the perceptual effects of slower modulations (up to 20Hz). Unit of fluctuation strength is vacil, which is defined as the fluctuation strength of a 1kHz tone of 60 dB amplitude modulated at 4Hz.

- PESQ; is a measure for assessing speech quality based on ITU 1116-1. PESQ can be considered as an objective estimation for speech signals. It was only applied to speech stimuli signals, with other types of contents like music and movie it does not return significant results.

3.6.2 Subjective Measures

Many subjective measures such as clarity, fullness, spaciousness, brightness, softness/-gentleness (in opposition to sharpness), absence of extraneous sounds and fidelity were derived in the very early sound quality analysis [46]. Here, we considered two criteria in designing the subjective measures. The first criterion is that the measures should be general enough to cover the perception in the TV industry. The second criterion is ease of describing the criterion to the jury without leaving room for ambiguity. Moreover, the criteria should be understandable by the general public and consumers as well. Thus, simple and common words were needed.

Considering the criteria above, for subjective evaluation of the TV samples, a total of four different scales were used: bass balance, treble balance, speech clarity, and overall sound quality. Before the subjective evaluation section, all participants took an orientation session where the subjective scales were described with technical details. Description of subjective metrics are given below;

- Bass balance is described as an emphasis on lower frequencies. Both under- and over-emphasis of low frequencies (less than 250Hz) are described as poor bass balance. Woofers or subwoofers are used in some of the higher-end TV systems to improve the bass balance. One of the TV set in our experimental setup has subwoofer as shown in table 2.

- Treble balance is associated with higher frequency content in the sound. A good representation of high-frequency content by a system makes sounds brighter and increases general pleasantness whereas poor representation generates muffled sound quality. Sometimes, tweeters are used in TVs to increase high-frequency content and improve treble balance. However, none of the TV sets used in this thesis contains tweeters.

- Speech quality is described as the naturalness of speech sounds. Listeners were

asked to score how natural the speech signals sound given a TV set.

- Overall sound quality is described as the average pleasantness of the audio (music, speech, movies) for a TV. Listeners were asked to judge the average pleasantness of the sounds by taking into account the other three scales.

CHAPTER IV

METHOD

4.1 Score Normalization

Score normalization is performed both on subjective and objective tests. For objective tests, the goal was to calibrate all scores in reference to the reference sample in the MUSHRA test. For subjective scores the goal was to increase consistency between different subjects.

For objective tests, scores are first averaged over all subjects and test samples and normalized by the average score of the reference sample, which gets the best score. Then, normalization is done by dividing the scores obtained from all stimuli signals by the score obtained for the reference signal. Hence, the objective test score for the reference signal was equal to 1 for all original signals.

Subjective test scores were normalized for each participant j, stimuli i, and t-TV (tv) using the following equation:

$$\tilde{V}_s^{i,j}(tv) = \frac{V_s^{i,j}(tv) - min(\mathbf{V_s^{(i,j)}}) + 1}{V_s^{(i,j)}(anc)}, \tag{1}$$

where $\tilde{V}_s^{i,j}(tv)$ represents the normalized subjective score, $\min(\mathbf{V_s^{(i,j)}})$ represents the minimum subjective score among all the t-TVs, $V_s(anc)$ is the score of the anchor signal and V_s represents the unnormalized subjective score in the range of 0 to 100. The minimum score for the test is subtracted from the other scores for compensating any constant bias that the participant might have. Because the scores are more meaningful relative to a baseline signal, they are divided by the anchor signal for further normalization.

4.2 Correlation of Objective and Subjective Scores

After normalization of objective and subjective scores, Pearson's correlation coefficient (PCC) was used to analyze the correlation between them. PCC was calculated as follows. For each subjective metric s, test stimuli i, and t-TV tv, the score vector

$$\tilde{V}_s^{(i)}(tv) = \frac{1}{N_p}\sum_{j=1}^{N_p}\tilde{V}_s^{(i,j)}(tv) \tag{2}$$

where j is the participant and N_p is the total number of participants. Then, the score vector for each t-TV $\tilde{\mathbf{V}}_\mathbf{s}(\mathbf{tv}) = [\tilde{V}_s^{(1)}(tv)\ \ \tilde{V}_s^{(1)}(tv)\ \ ...\ \ \tilde{V}_s^{(N_t)}(tv)]$ where N_t is the total number of test stimuli.

Similarly, the objective score for each stimuli i and t-TV is

$$\tilde{V}_o^{(i)}(tv) = \frac{1}{N_p}\sum_{j=1}^{N_p}\tilde{V}_o^{(i,j)}(tv), \tag{3}$$

and $\tilde{\mathbf{V}}_\mathbf{o}(\mathbf{tv}) = [\tilde{V}_o^{(1)}(tv)\ \ \tilde{V}_o^{(1)}(tv)\ \ ...\ \ \tilde{V}_o^{(N_t)}(tv)]$.

For each test, correlation between the subjective scores $\tilde{\mathbf{V}}_\mathbf{s} = [\tilde{\mathbf{V}}_\mathbf{s}(1)\ ...\ \tilde{\mathbf{V}}_\mathbf{s}(7)]$ and the objective scores $\tilde{\mathbf{V}}_\mathbf{o} = [\tilde{\mathbf{V}}_\mathbf{o}(1)\ ...\ \tilde{\mathbf{V}}_\mathbf{o}(7)]$ is computed using Pearson's correlation

$$\rho(\tilde{\mathbf{V}}_\mathbf{o}, \tilde{\mathbf{V}}_\mathbf{s}) = \frac{Cov(\tilde{\mathbf{V}}_\mathbf{o}, \tilde{\mathbf{V}}_\mathbf{s})}{\sigma_{\tilde{\mathbf{V}}_\mathbf{s}}, \sigma_{\tilde{\mathbf{V}}_\mathbf{o}}} \tag{4}$$

Correlation is assumed to be strong if its absolute value is greater than 0.5; and, it is assumed to be significant if the significance of the null hypothesis is less than 0.05.

CHAPTER V

RESULTS AND DISCUSSION

5.1 Subjective Results

In order to investigate the factors that affect the subjective scores, impulse responses of the speakers can be compared from Figure 5 to Figure 10. Loudspeakers in t-TV 4-5-6 have better low-frequency and roll-off are compared to t-TV 1-2-3. Roll-off frequency of TV-5 is lower than TV-6 because it has a subwoofer. As a result, t-TV 4-5-6 can render lower frequencies (below 150Hz) better. However, there is typically no speech formant at those low frequencies and rendering them better does not significantly affect the speech quality. The high-frequency spectrum of the cheaper loudspeakers in TV 1-2-3 is significantly better compared to TV 4-5-6; and, rendering higher frequency content better resulted in higher speech quality score for TV-2.

5.1.1 Bass Balance

Normalized subjective scores for bass balance is shown in figure 14. Since t-TV 4-5-6 all have the same loudspeaker model, C, which is a relatively more expensive loudspeaker, results for these TV samples are very similar and better than t-TV 1-2-3. Unlike t-TV 4-5 the t-TV 6 has a subwoofer which enhances the rendering of low-frequency content, thus it has better subjective results than the other two TVs. T-TV 4 has slightly better performance than t-TV 5 since it has a high-end audio amplifier. Thus it can be considered that audio power amplifier has a small effect on listeners bass balance scores. Because loudspeakers in t-TV 1-2-3 do not have an enclosure, they have poor performance on bass balance evaluation. The outcome from bass balance evaluation is TV samples which have loudspeakers with enclosure provides relatively good bass balance results.

5.1.2 Treble Balance

Treble balance is related to the emphasis on high-frequency content, Normalized subjective scores for treble balance is shown in Figure 15. Unlike bass balance, the results are almost opposite for the treble balance where the listeners had a higher preference for TV 1-2-3. Due to t-TV 1 has high-end audio power amplifier, it has better results than t-TV 2-3. The one can easily notice from Figure 15 that there is a significant difference between t-TV 1-2-3 and t-TV 4-5-6. That much of difference between this TV samples on treble balance score is not expected. The major difference between loudspeaker models A-B and C is the enclosure. Loudspeakers with enclosure have better bass performance, and similar treble performance, see impulse responses from Figure 5 to Figure 10. Thus, it is expected to have similar treble balance scores for all the TV samples. Due to the masking effect [47] between low and high-frequency component it does not result as expected. Tv samples, which have better low-frequency roll-off, have poor treble balance. Increasing the low-frequency gain with a subwoofer provides the worst treble balance score, see results for t-TV 5 in figure 15. Therefore, there is a trade-off between bass and treble balances in listeners judgments. To balance this trade-off, 3-way driven loudspeakers (mid-range, woofer, tweeter) might be a better solution.

5.1.3 Speech Quality

Normalized subjective scores for speech quality is shown in figure 16. For speech quality, listeners preferred TVs that have better treble balance. In fact, the treble balance results are almost the same as the speech quality results. High-frequency content improves the brightness of speech, and emphasis of that appears to have a significant impact on the listeners' subjective preference even when the lower frequency content is rendered better.

5.1.4 Overall Sound Quality

Interestingly, even though the TVs with cheaper sets have better scores in treble and speech quality tests, the more expensive sets performed better in overall sound quality tests. Thus, the bass balance was a good predictor of the overall sound quality in the TV context. These results suggest that for the speech signal emphasis should be more in high-frequency content whereas, for other sounds, such as music, bass balance is more important. For music, though, frequencies below 200 Hz are very important for some instruments such as drums. Similarly, most real-life noise signals in movies, such as babble or car noises, are low-frequency signals. For those cases, t-TV 4-5-6 expected to perform better, which resulted in a higher overall score for TVs with lower cut-off frequencies. Normalized subjective scores for overall sound quality is shown in figure 17.

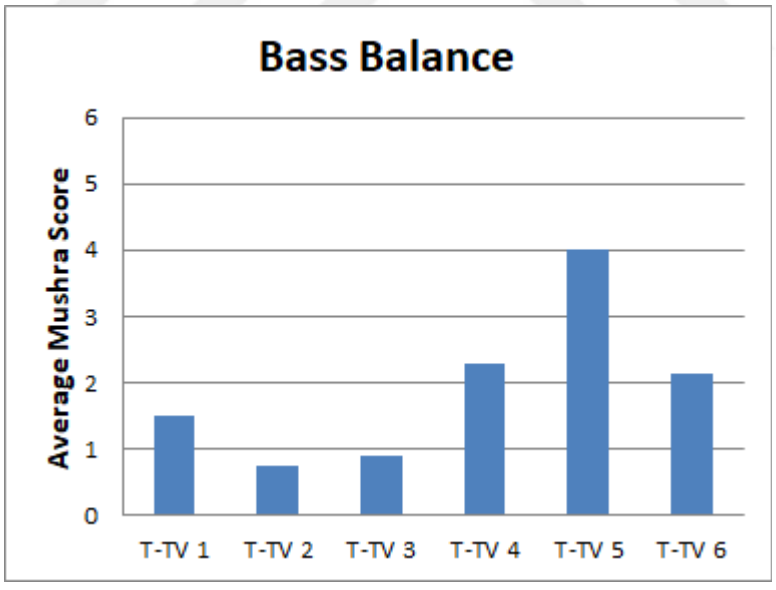

Figure 14: Normalized MUSHRA scores for Bass Balance

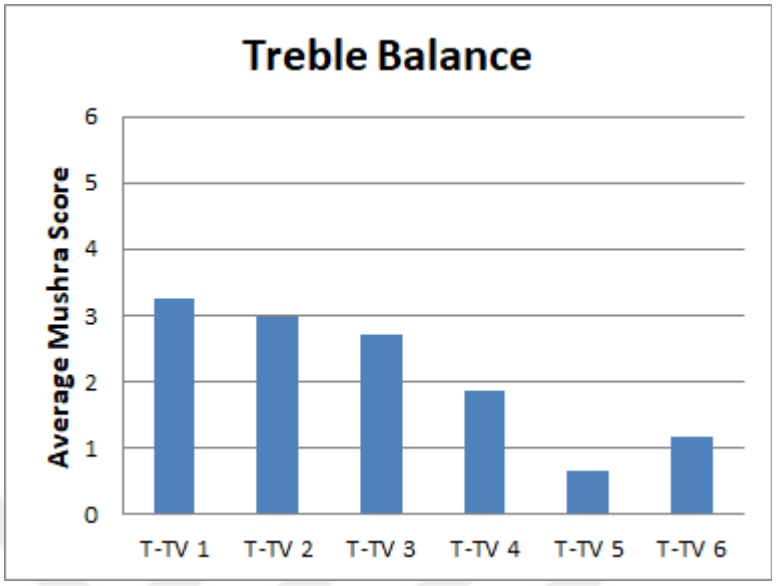

Figure 15: Normalized MUSHRA scores for Treble Balance

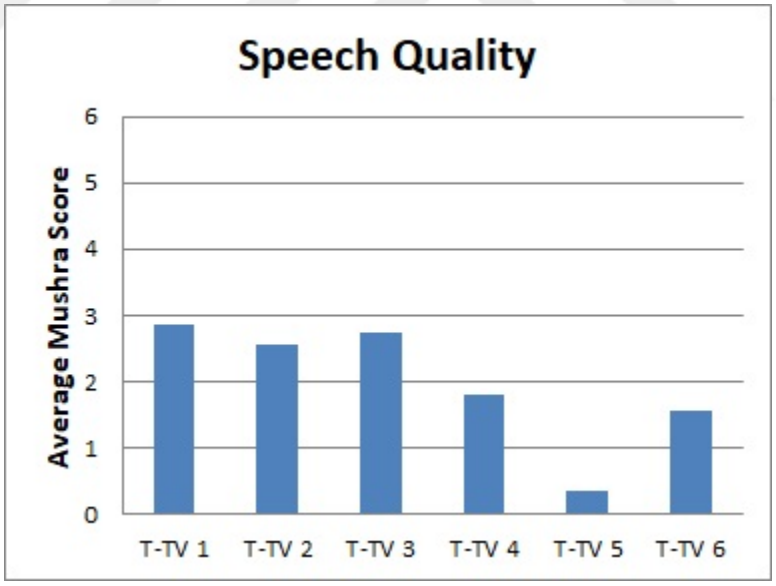

Figure 16: Normalized MUSHRA scores for Speech Quality

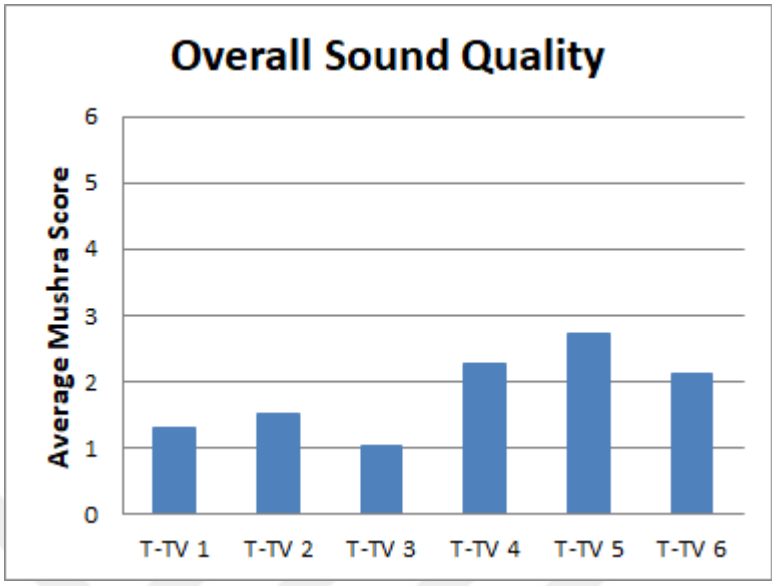

Figure 17: Normalized MUSHRA scores for Overall Sound Quality

5.2 Correlation Analysis

Correlations of objective and subjective scores are shown in Table 7 and Table 8. Sharpness measure correlated strongly with speech quality, treble and bass balance. Correlation is positive for the treble balance and speech quality, and it is negative for the bass balance. This result is consistent with the subjective test results where loudspeakers with better high-frequency response had higher treble balance and speech quality, but lower bass balance. Sharpness, which is a measure of high-frequency content to low-frequency content ratio, strongly correlated with that finding in the subjective tests. Thus, sharpness measure can be used, for example, for assessing if a subwoofer in the loudspeaker has a negative effect on rendered speech quality or if a tweeter is needed.

Roughness, which is an indicator of how well the high-frequency amplitude modulations ($> 15Hz$) are rendered by the loudspeaker, was not found to strongly correlate with any of the subjective test results. However, fluctuation strength, which indicates the rendering accuracy for lower frequency ($< 20Hz$) amplitude modulations,

correlated strongly with treble balance and speech quality. Thus, more accurate rendering of longer-range syllable-level modulations, captured with fluctuation strength, are better correlated with the perceived speech quality as opposed to short-duration modulations that are captured with the roughness measure.

Loudness was found to be significantly correlated with treble balance as shown in Figure 19. Results for the three samples (jazz, techno, electro music) are clustered together for each TV in Figure 19, which indicates that the loudness-treble balance relationship is not very sensitive to the test samples. Loudspeakers with better treble balance are perceived louder for music because the human ear is more sensitive to air pressure at higher frequencies than lower frequencies. Musical instruments can generate high-energy high-frequency sounds that are rendered well with the t-TVs with better treble balance and loudness successfully captured that effect.

Even though louder t-TVs subjectively had better treble balance, they did not have better-perceived speech quality. Because most of the energy in the speech signal is below 2kHz, TV-5 and TV-6, which have worse speech quality but better bass balance, can render louder sounds as shown in figure 20. That results in a slightly negative correlation of loudness with the speech quality. Still, that correlation is weak and statistically not significant.

Correlation of PESQ scores and sharpness scores with the subjective test results are very similar. Sharpness correlates slightly better with the speech quality and it has a stronger correlation with the treble and bass balances. Moreover, samples from the same t-TVs are clustered better with the sharpness feature as shown in figure 18 and figure 21 and, that indicates better score invariance with regards to the test samples. Based on these findings, sharpness can be used as a better alternative to the widely-used PESQ measure for the assessment of speech quality in TVs.

Table 7: Correlation scores between the subjective and objective metrics.

Objective Metrics	Overall Sound Quality	Bass Balance	Treble Balance	Speech Quality
Sharpness	0.378	**-0.782**	**0.814**	**0.764**
Roughness	0.173	-0.129	0.167	0.287
Loudness	0.131	-0.025	**0.834**	-0.273
Fluctuation	**0.495**	**-0.490**	**0.555**	**0.531**
PESQ	0.176	**-0.579**	**0.659**	**0.759**

Table 8: Significance scores between the subjective and objective metrics.

Objective Metrics	Overall Sound Quality	Bass Balance	Treble Balance	Speech Quality
Sharpness	6×10^{-2}	2×10^{-5}	2×10^{-6}	4×10^{-9}
Roughness	9×10^{-1}	3×10^{-1}	4×10^{-1}	7×10^{-2}
Loudness	4×10^{-1}	9×10^{-1}	2×10^{-6}	1×10^{-1}
Fluctuation	7×10^{-3}	2×10^{-5}	9×10^{-3}	3×10^{-5}
PESQ	2×10^{-2}	6×10^{-5}	2×10^{-6}	6×10^{-9}

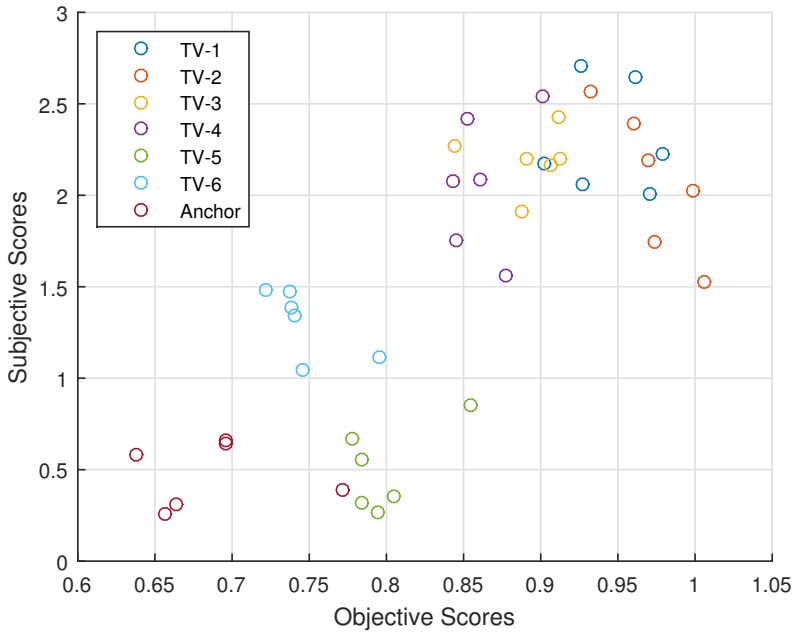

Figure 18: Scatter diagram of normalized subjective (Speech Quality) and objective scores (Sharpness).

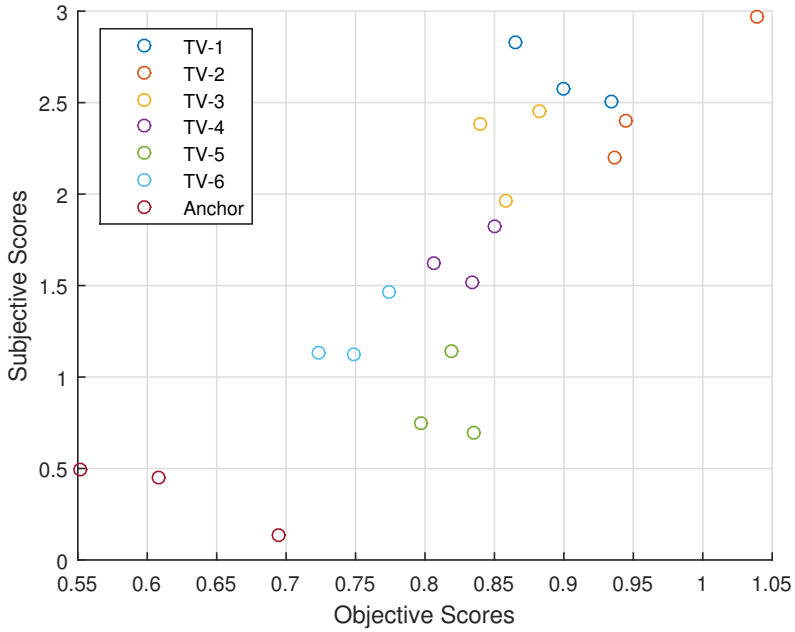

Figure 19: Scatter diagram of normalized subjective (Treble Balance) and objective (Loudness) scores.

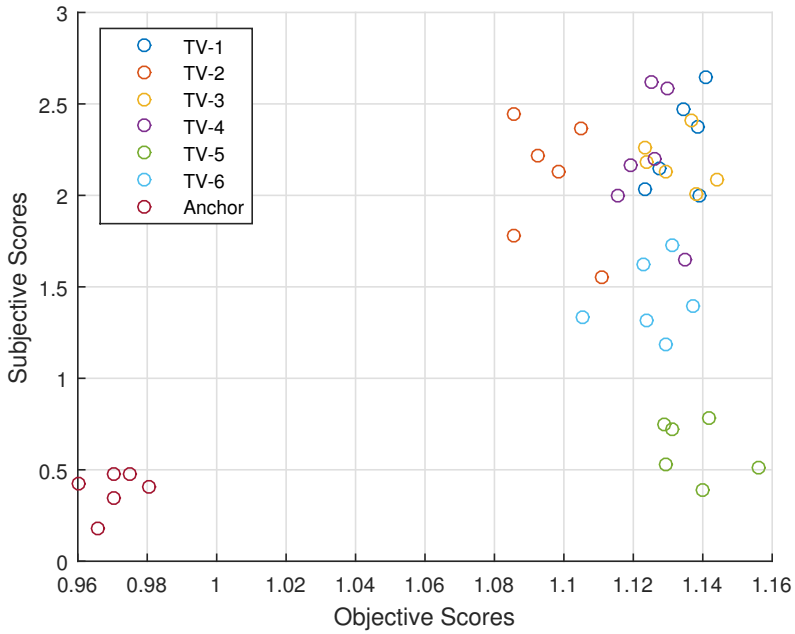

Figure 20: Scatter diagram of normalized subjective (Speech Quality) and objective (Loudness) scores.

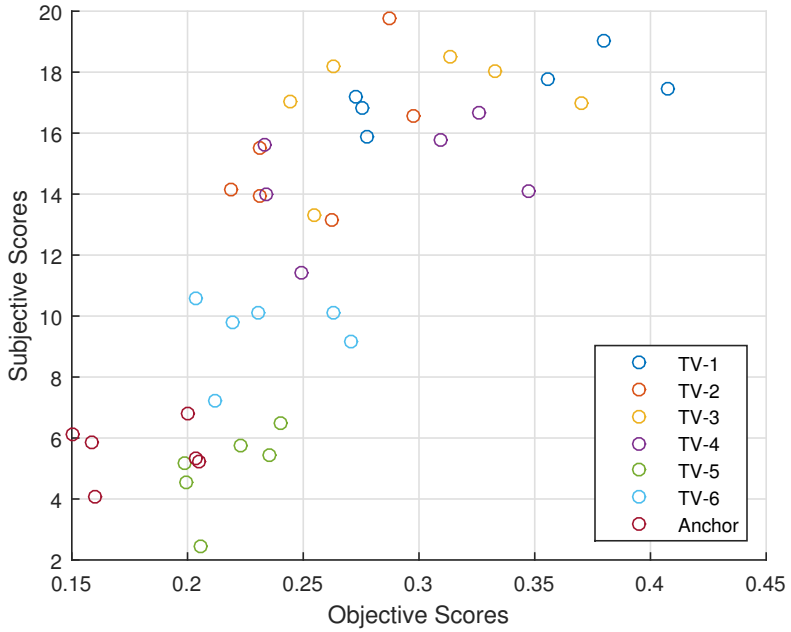

Figure 21: Scatter diagram of normalized subjective (Speech Quality) and objective (PESQ) scores.

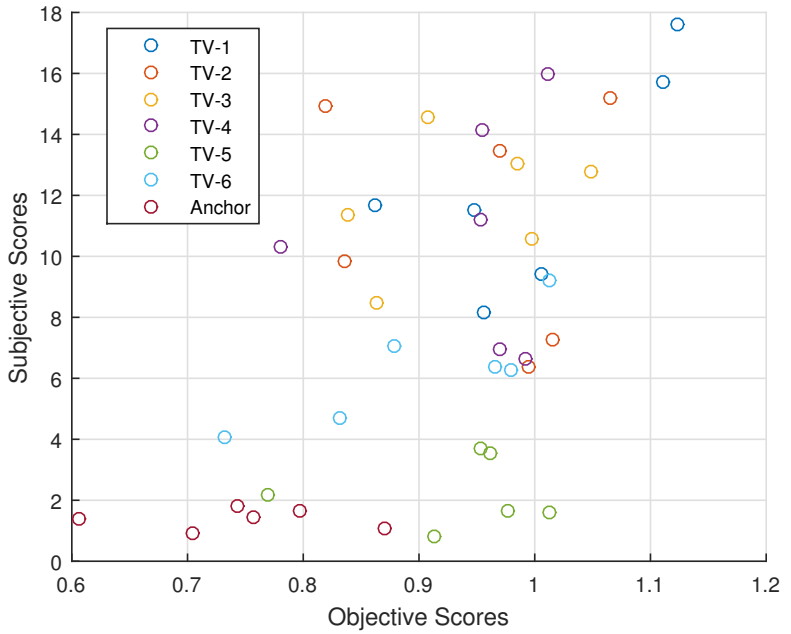

Figure 22: The linear regression line for normalized subjective (Speech Quality) and objective (Fluctuation Strength) scores.

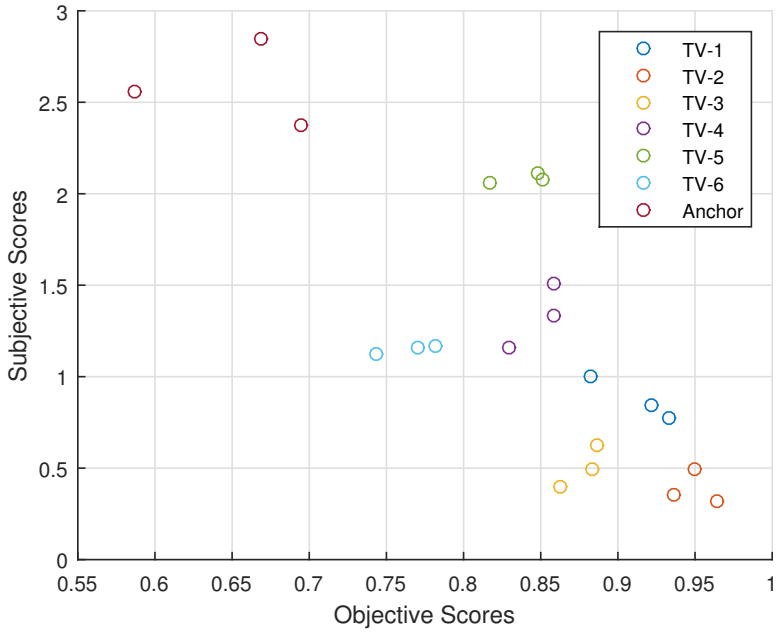

Figure 23: The linear regression line for normalized subjective (Bass Balance) and objective (Sharpness) scores.

CHAPTER VI

CONCLUSION

This thesis presents subjective and objective test guidelines that can be used for TV manufacturers for sound quality assessment together with an analysis of the correlation between some of the commonly-used objective and subjective measures. Tests were done with 6 different TV samples and 24 listeners. Listeners subjectively evaluated music samples, sound effects in movies, and speech samples using the MUSHRA test. Four subjective measures were used to evaluate the subjective assessment of TV sound quality. Correlation analysis between objective and subjective metrics was done.

We concluded with the following as a result of the correlation analysis:

- Sharpness is a good indicator of bass balance and treble balance. It also correlates well with the perceived speech quality.
- Roughness measure does not correlate with any of the subjective measures.
- Commonly used loudness measure correlates well with the treble balance but the correlation is not significantly higher than the correlation of sharpness and treble balance.
- Fluctuation strength correlates well with all measures and it is the only measure that has a strong correlation with the overall sound quality.
- PESQ measure, which is commonly used for assessing the quality of speech signals, had a strong correlation with the speech quality. However, the correlation was not significantly better than the correlation of sharpness with speech quality.

APPENDIX A

MATLAB CODES

Correlation analysis between objective and subjective metrics are computed with the script below which is a sample script to compute the correlation between speech quality and fluctuation strength. One can compute the correlation coefficient between any objective and subjective metrics by modifying this script. The input of the m-file is subjective and objective result vectors. Outputs are correlation coefficient, significance scores and scatter diagrams.

```
clc
clear

% Load subjective and objective results.
load('subj_Speech_Quality.mat')
load('obj_FlucStrenght')


% Remove bad subjects
% Reference score should be 100 or close to 100
% Anchor score should be very low and below others.
bad_subj = [5 8 9 16 17 18 19];
for i = 1:length(bad_subj)
    subj_Speech_Quality(:, (bad_subj-1)*6+1:bad_subj*6) = [];
end

```

```
% Subjective score normalization
subj_norm = subj_Speech_Quality;
for p = 1:102
    tmp = subj_Speech_Quality(:,p);
    subj_norm(:, p) = 6*(tmp - min(tmp)) / tmp(7);
end

% Objective score normalization
obj_norm = obj_FlucStrenght;
for s = 1:6
    tmp = obj_FlucStrenght(:,s);
    obj_norm(:, s) = tmp / tmp(end);
end

% Calculate correlation coefficients and significancy
figure;
col = ['ro';'bo';'ko';'yo';'mo';'go']
res = zeros(2, 6*6);
c = 1;
for t = 1:7
    avg_subj = zeros(6, 6);
    for s = 1:6
        avg_subj(t,s) = mean( subj_norm(t, s:6:end ) );
        res(1,c) = avg_subj(t,s);
        res(2,c) = mean(obj_norm(t,s));
        c = c+1;
    end
```

```
44 end
45
46 [R,P] = corrcoef(res(2,:), res(1,:))
47 ['correlation is : ' num2str(R(2,1))]
48 ['significance is : ' num2str(P(2,1))]
49
50 % Sketch the figure
51 figure;
52 for i = 1:7
53     hold on;scatter(res(2,1+6*(i-1):6*i), res(1,1+6*(i-1):6*i
       ));
54 end
55 grid on; legend('TV-1', 'TV-2','TV-3','TV-4', 'TV-5','TV-6','
   Anchor')
```

APPENDIX B

OBJECTIVE RESULTS

The objective test results which are strongly correlated with subjective metrics are shown in Table 9 to Table 26.

Table 9: Fluctuation Strength(vacil) results with speech quality stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Male 1	2.385	2.479	2.152	2.419	2.276	2.074	1.886	2.493
Male 2	1.808	1.752	1.757	1.636	1.613	1.535	1.477	2.097
Male 3	1.181	1.192	1.171	1.164	1.146	1.133	0.9355	1.174
Female 1	1.5	1.296	1.558	1.508	1.509	1.39	0.9591	1.582
Female 2	2.901	2.532	2.371	2.493	2.511	2.559	2.271	2.611
Female 3	1.968	1.867	1.836	1.772	1.775	1.773	1.302	1.752

Table 10: PESQ results with speech quality stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Male 1	2.132	1.572	1.8811	1.855	1.336	1.384	1.2	5
Male 2	1.637	1.387	1.4665	1.401	1.1924	1.3148	0.9	5
Male 3	1.6511	1.3861	1.578	1.4955	1.1956	1.2195	0.95	5
Female 1	2.4462	1.725	2.2225	2.0839	1.4122	1.5792	1.22	5
Female 2	2.2772	1.7856	1.9984	1.9575	1.4397	1.623	1.23	5
Female 3	1.6633	1.3145	1.53	1.4033	1.233	1.2721	0.96	5

Table 11: Sharpness(acum) results with speech quality stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Male 1	1.853	2.001	1.862	1.736	1.653	1.522	1.31	2.054
Male 2	1.594	1.627	1.452	1.402	1.278	1.214	1.07	1.629
Male 3	1.563	1.62	1.47	1.412	1.263	1.189	1.12	1.61
Female 1	1.729	1.738	1.574	1.571	1.45	1.345	1.297	1.864
Female 2	1.578	1.593	1.497	1.48	1.403	1.306	1.267	1.642
Female 3	1.525	1.58	1.462	1.403	1.308	1.214	1.093	1.646

Table 12: Loudness(phon) results with speech quality stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Male 1	84.67	81.5	84.34	83.78	84.79	83.01	72.11	75.09
Male 2	80.59	78.04	80.23	79.51	81.12	80.25	68.62	71.05
Male 3	78.09	76.14	78.43	77.8	79.24	77.97	66.52	68.55
Female 1	86.11	83.2	86.14	86.32	86.52	86.06	74.72	76.65
Female 2	86.06	82.6	85.92	85.07	85.52	84.96	74.11	75.6
Female 3	81.87	79.28	81.67	81.05	81.78	81.15	69.64	71.75

Table 13: Sharpness(acum) results with bass balance stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Jazz	1.669	1.856	1.576	1.519	1.492	1.382	1.241	1.786
Pop	2.015	2.198	1.956	1.878	1.856	1.683	1.285	2.328
Electro	1.953	2.035	1.864	1.812	1.779	1.627	1.321	2.172

Table 14: Sharpness(acum) results with treble balance stimuli signals

Audio Samples	t-TV 1	t-TV 2	t-TV 3	t-TV 4	t-TV 5	t-TV 6	Anchor	Reference
Pop	2.076	2.202	2.03	1.951	1.921	1.749	1.38	2.352
Rock	1.816	1.899	1.747	1.692	1.676	1.518	1.317	1.97
Hip-Hop	1.777	1.808	1.682	1.635	1.614	1.488	1.322	1.904

APPENDIX C

SCATTER DIAGRAM DATA POINTS

Table 15: X-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective scores (Sharpness)

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	0.902	0.9785	0.970	0.927	0.961	0.926
t-TV 2	0.974	0.998	1.006	0.932	0.970	0.959
t-TV 3	0.906	0.891	0.913	0.844	0.911	0.888
t-TV 4	0.845	0.860	0.877	0.842	0.901	0.852
t-TV 5	0.804	0.784	0.784	0.777	0.854	0.794
t-TV 6	0.740	0.745	0.738	0.721	0.795	0.737
Anchor	0.637	0.656	0.695	0.695	0.771	0.664

Table 16: Y-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective scores (Sharpness)

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	2.173	2.228	2.003	2.060	2.642	2.711
t-TV 2	1.747	2.029	1.527	2.571	2.194	2.390
t-TV 3	2.164	2.202	2.199	2.268	2.423	1.914
t-TV 4	1.751	2.087	1.565	2.077	2.543	2.419
t-TV 5	0.353	0.559	0.318	0.671	0.856	0.267
t-TV 6	1.344	1.047	1.386	1.486	1.115	1.470
Anchor	0.578	0.255	0.664	0.643	0.388	0.307

Table 17: X-axis data set for Scatter diagram of normalized subjective (Treble Balance) and objective(Loudness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3
t-TV 1	0.934	0.866	0.899
t-TV 2	1.039	0.944	0.937
t-TV 3	0.882	0.840	0.858
t-TV 4	0.851	0.807	0.834
t-TV 5	0.835	0.797	0.819
t-TV 6	0.774	0.723	0.749
Anchor	0.695	0.552	0.608

Table 18: Y-axis data set for Scatter diagram of normalized subjective (Treble Balance) and objective(Loudness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3
t-TV 1	2.509	2.831	2.577
t-TV 2	2.967	2.401	2.201
t-TV 3	2.451	2.380	1.962
t-TV 4	1.821	1.626	1.517
t-TV 5	0.693	0.750	1.146
t-TV 6	1.466	1.133	1.127
Anchor	0.133	0.493	0.447

Table 19: X-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(Loudness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	1.128	1.134	1.139	1.123	1.138	1.141
t-TV 2	1.085	1.098	1.111	1.085	1.093	1.105
t-TV 3	1.123	1.129	1.144	1.124	1.137	1.138
t-TV 4	1.116	1.119	1.135	1.126	1.125	1.130
t-TV 5	1.129	1.142	1.156	1.129	1.131	1.140
t-TV 6	1.105	1.129	1.137	1.123	1.124	1.131
Anchor	0.960	0.966	0.970	0.975	0.980	0.971

Table 20: Y-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(Loudness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	2.173	2.228	2.003	2.060	2.642	2.711
t-TV 2	1.747	2.029	1.527	2.571	2.194	2.390
t-TV 3	2.164	2.202	2.199	2.268	2.423	1.914
t-TV 4	1.751	2.087	1.565	2.077	2.543	2.419
t-TV 5	0.353	0.559	0.318	0.671	0.856	0.267
t-TV 6	1.344	1.047	1.386	1.486	1.115	1.470
Anchor	0.578	0.255	0.664	0.643	0.388	0.307

Table 21: X-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(PESQ) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	0.355	0.273	0.275	0.408	0.380	0.277
t-TV 2	0.262	0.231	0.231	0.288	0.298	0.219
t-TV 3	0.314	0.244	0.263	0.370	0.333	0.255
t-TV 4	0.309	0.234	0.249	0.347	0.326	0.234
t-TV 5	0.223	0.199	0.199	0.235	0.240	0.206
t-TV 6	0.231	0.219	0.203	0.263	0.271	0.212
Anchor	0.200	0.150	0.158	0.203	0.205	0.160

Table 22: Y-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(PESQ) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	17.793	17.216	16.845	17.434	19.013	15.881
t-TV 2	13.134	15.501	13.936	19.776	16.583	14.147
t-TV 3	18.481	17.058	18.195	16.975	18.024	13.288
t-TV 4	15.784	15.602	11.396	14.108	16.653	14.009
t-TV 5	5.727	5.161	4.539	5.444	6.505	2.420
t-TV 6	10.098	9.778	10.565	10.119	9.146	7.227
Anchor	6.795	6.122	5.836	5.347	5.239	4.057

Table 23: X-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(Fluctuation Strength) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	0.957	0.862	1.006	0.948	1.111	1.123
t-TV 2	0.994	0.835	1.015	0.819	0.970	1.066
t-TV 3	0.863	0.838	0.997	0.985	0.908	1.048
t-TV 4	0.970	0.780	0.991	0.953	0.955	1.011
t-TV 5	0.913	0.769	0.976	0.954	0.962	1.013
t-TV 6	0.832	0.732	0.965	0.879	0.980	1.012
Anchor	0.757	0.704	0.797	0.606	0.870	0.743

Table 24: Y-axis data set for Scatter diagram of normalized subjective (Speech Quality) and objective(Fluctuation Strength) scores

t-TV Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
t-TV 1	8.176	11.662	9.412	11.498	15.737	17.619
t-TV 2	6.378	9.839	7.257	14.909	13.480	15.205
t-TV 3	8.500	11.340	10.594	13.037	14.555	12.763
t-TV 4	6.979	10.293	6.655	11.225	14.122	15.962
t-TV 5	0.831	2.179	1.662	3.702	3.539	1.589
t-TV 6	4.717	4.056	6.351	7.059	6.249	9.198
Anchor	1.463	0.929	1.663	1.383	1.056	1.793

Table 25: X-axis data set for Scatter diagram of normalized subjective (Bass Balance) and objective(Sharpness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3
t-TV 1	0.883	0.922	0.933
t-TV 1	0.936	0.964	0.950
t-TV 1	0.863	0.887	0.883
t-TV 1	0.830	0.859	0.859
t-TV 1	0.817	0.851	0.848
t-TV 1	0.744	0.771	0.782
Anchor	0.587	0.669	0.694

Table 26: Y-axis data set for Scatter diagram of normalized subjective (Bass Balance) and objective(Sharpness) scores

t-TV Samples	Sample 1	Sample 2	Sample 3
t-TV 1	0.998	0.845	0.774
t-TV 2	0.355	0.318	0.493
t-TV 3	0.394	0.630	0.491
t-TV 4	1.156	1.508	1.330
t-TV 5	2.062	2.078	2.111
t-TV 6	1.125	1.161	1.167
Anchor	2.560	2.849	2.373

Bibliography

[1] Bruel and Kjaer, "Figure captured from training documents.".

[2] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Munich, GE: Springer-Verlag Berlin Heidelberg, 3 ed., 2006.

[3] T. I. R. Assembly, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound system," no. Rec. ITU-R BS.1116-1, 1994-1997.

[4] T. I. R. Assembly, "Method for the subjective assessment of intermediate quality level of coding systems," no. Rec. ITU-R BS.1534-1, 2001-2003.

[5] D. Bowen, "Correlating sound quality metrics and jury ratings," *Sound And Vibration*, September, 2008.

[6] D. S. S. P. Y. Ih, J.G. Lim, "Experimental design and assessment of product sound quality: Application to a vacuum cleaner," *Noise Control Engineering Journal*, vol. 51, no. 4, pp. 244,252, July 2013.

[7] A. M. Willemsen and R. D., "Characterization of sound quality of impulsive sounds using loudness based metric," *International Congress on Acoustics*, vol. 20, Agust 2010.

[8] J. Blauert, *An introduction to binaural technology*. Berlin, GE: Springer-Verlag Berlin Heidelberg, 1 ed., 1997.

[9] T. Walton, "The overall listening experience of binaural audio," 09 2017.

[10] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 01 2000.

[11] ITU, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 01 2001.

[12] J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part ii-perceptual model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[13] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 408–419, 2003.

[14] B. Moore, C.-T. Tan, N. Zacharov, and V.-V. Mattila, "Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion," *Journal of the Audio Engineering Society*, vol. 52, pp. 1228–1244, 12 2004.

[15] F. Toole, "The measurement and calibration of sound reproducing systems," *J. Audio Eng. Soc*, vol. 63, no. 7/8, pp. 512–541, 2015.

[16] S. Bech, "Listening tests on loudspeakers: A discussion of experimental procedures and evaluation of the response data," in *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*, May 1990.

[17] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences: Part 1," *J. Audio Eng. Soc*, vol. 34, no. 4, pp. 227–235, 1986.

[18] B. B. N. Porter, "A study of standard methods for measuring the sound quality of industrial products: Final report," *National Physical Laboratory*, pp. 11,12, 1997.

[19] S. Kraft and U. Zolzer, "Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality," *Linux Audio Conference*, p. 102;109, 2014.

[20] M. Bodden and R. Heinrichs, "Diesel sound quality analysis and evaluation," 07 2019.

[21] E. Parizet, E. Guyader, and V. Nosulenko, "Analysis of car door closing sound quality," *Applied Acoustics*, vol. 69, pp. 12–22, 01 2008.

[22] D. L. Bowen, "Sound quality studies of front-loading washing machines," *Sound and vibration*, vol. 44, pp. 8–13, 12 2010.

[23] B. Akhmetov, S. Gupta, and K. Ahuja, "Noise source ranking of a hairdryer," 06 2014.

[24] S. FUJIWARA, D. SAKAI, A. IWAHARA, and T. Toi, "Sound design of vacuum cleaner based on sound quality evaluation," *The Proceedings of the Symposium on Environmental Engineering*, vol. 2003.13, pp. 71–72, 06 2003.

[25] A. Pras, R. Zimmerman, D. Levitin, and C. Guastavino, "Subjective evaluation of mp3 compression for different musical genres," p. 9–12, 10 2009.

[26] A. Pras and C. Guastavino, "Sampling rate discrimination: 44.1 khz vs. 88.2 khz," vol. 2, 05 2010.

[27] C. E. Association and A. N. S. Institute, *CEA Standard: Test Methods of Measurement for Audio Amplifiers : CEA-490-A R-2008*. Consumer Electronics Association, Technology & Standards Department, 2008.

[28] I. I. E. Commission, *IEC Standard: Sound system equipment – Part 5: Loudspeakers : IEC 60268-5.* IEC - International Electrotechnical Commission, 2007.

[29] G. Stoll and F. Kozamernik, "Ebu listening tests on internet audio codecs," *EBU Technical Review*, 08 2000.

[30] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences: Part 1," *J. Audio Eng. Soc*, vol. 34, no. 4, pp. 227–235, 1986.

[31] E. S. M. T. Inc., "Digital input high power class d audio amplifier, product id:ad82586c.".

[32] T. Instruments, "10-w/15-w digital audio power amplifier with integrated cap-free hp amplifier, product id:tas5719.".

[33] A. Precision, "Apx52x b series audio analyzers," 2015.

[34] Earthworks, "30khz measurement microphone, model id: M30.".

[35] M-Audio, "Two-channel usb 2.0 audio interface with 24-bit/96 khz resolution, model id: M-track plus ii.".

[36] R. E. Wizard, "free room acoustics analysis software for measuring and analyzing room and loudspeaker responses.".

[37] V. Koehl and M. Paquier, "Loudspeaker sound quality: comparison of assessment procedures," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3298–3298, 2008.

[38] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, "Audio quality evaluation by experienced and inexperienced listeners," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 060016, 2013.

[39] S. Olive, "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study," *AES: Journal of the Audio Engineering Society*, vol. 51, pp. 806–825, 09 2003.

[40] E. M. Wenzel and S. H. Foster, "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 102–105, Oct 1993.

[41] D. S. Brungart and G. D. Romigh, "Spectral hrtf enhancement for improved vertical-polar auditory localization," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 305–308, Oct 2009.

[42] "Bruel - kjaer binarual in-ear microphone set, type 4101-b, https://www.bksv.com/-/media/literature/product-data/bp2562.ashx.".

[43] "Bruel - kjaer time data recorder type-7708, https://www.bksv.com/-/media/literature/product-data/bp0005.ashx.".

[44] "Stax electronics audio products, stax sr-l700mk2, https://staxaudio.com/earspeaker/sr-l700.".

[45] Bruel and Kjaer, "Pulse 21, sound quality measurement software," 2016.

[46] A. Gabrielsson and H. Sjogren, "Perceived sound quality of sound-reproducing systems," *The Journal of the Acoustical Society of America*, vol. 65, pp. 1019–33, 05 1979.

[47] H. Patra, C. Roup, and L. Feth, "Masking of low-frequency signals by high-frequency, high-level narrow bands of noise," *The Journal of the Acoustical Society of America*, vol. 129, pp. 876–87, 02 2011.

VITA

Çağlar İşlek is a graduate student of electronics engineering at Özyeğin University, Istanbul, Turkey. He received his license degree in electrical and electronics engineering at Dokuz Eylül University in 2015. After he received his undergraduate degree in 2015, he started to work as a test engineer at Vestel Elektronik, Manisa, Turkey. He has been working on the testing of acoustical performance of Television products for 4 years. He has been working on the acoustical performance of smart voice assistant devices recently.