

2017

# Towards improving ViSQOL (Virtual Speech Quality Objective Listener) Using Machine Learning Techniques

Joseph McNally

*Technological University Dublin*

# Towards improving ViSQOL (Virtual Speech Quality Objective Listener) Using Machine Learning Techniques


Joseph McNally

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

M.Sc. in Computing (Advanced Software Development)

July, 2017

## Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Advanced Software Development), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University. The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

I hold no positions, short or long, in any of the firms included in this study, nor was the work sponsored by any third-party entity. Further, this research is not intended to be construed as providing or implying investment advice.

**Signed:** _____

**Joseph McNally**

**Date:** **31st July 2017**

## Abstract

Vast amounts of sound data are transmitted every second over digital networks. VoIP services and cellular networks transmit speech data in increasingly greater volumes. Objective sound quality models provide an essential function to measure the quality of this data in real-time. However, these models can suffer from a lack of accuracy with various degradations over networks. This research uses machine learning techniques to create one support vector regression and three neural network mapping models for use with ViSQOLAudio. Each of the mapping models (including ViSQOL and ViSQOLAudio) are tested against two separate speech datasets in order to comparatively study accuracy results. Despite the slight cost in positive linear correlation and slight increase in error rate, the study finds that a neural network mapping model with ViSQOLAudio provides the highest levels of accuracy in objective speech quality measurement. In some cases, the accuracy levels can be over double that of ViSQOL. The research demonstrates that ViSQOLAudio can be altered to provide an objective speech quality metric greater than that of ViSQOL.

**Key Words**: Objective Sound Quality, ViSQOL, ViSQOLAudio, Support Vector Regression, Neural Networks

## Acknowledgements

I would like to express my thanks to **Thibaut Lust** for providing essential guidance throughout every step of the creation of this dissertation. His encouragement, insight, experience and patience have been invaluable in helping me achieve my goals with this research.

Additional thanks must go to **Dr. Andrew Hines** for helping shape the proposal for this research, as well as providing valuable information and materials when needed.

Lastly, and most importantly, I wish to express my sincere and eternal gratitude to my fiancée **Rachel Hamilton**. Without her unending love and support, I would never have had the strength to push myself to greater academic heights with this research.

# Table of Contents

## Table of Figures

## Table of Tables

## List of Acronyms

**ACR** - Absolute Category Rating

**ADSL** - Asymmetrical Digital Subscriber Line

**CCR** - Comparative Category Rating

**DCR** - Degradation Category Rating

**DNN** - Deep Neural Network

**DRT** - Diagnostic Rhyme Test

**FTTH** - Fibre-To-The-Home

**fwSNRseg** - Frequency-Weighted Segmental SNR

**GPU** - Graphical Processing Unit

**GSM** - Global System for Mobile Communications

**IETF** - Internet Engineering Task Force

**IP** - Internet Protocol

**ISDN** - Integrated Services Digital Network

**ITU** - International Telecommunication Union

**LLR** - Log-likelihood ratio

**MOS** - Mean Opinion Score

**MOS-LQO** - Mean Opinion Score Listening Quality Objective

**MOS-LQS** - Mean Opinion Score Listening Quality Subjective

**MRT** - Modified Rhyme Test

**NSIM** - Neurogram Similarity Index Measure

**PEAQ** - Perceptual Evaluation of Audio Quality

**PESQ** - Perceptual Evaluation of Speech Quality

**POLQA** - Perceptual Objective Listening Quality Assessment

**POTS** - Plain Old Telephony Services

**QoE** - Quality of Experience

**QoS** - Quality of Service

**RMSE** - Root-mean-square-error

**SNR** - Signal-To-Noise Ratio

**SOS** - Standard deviation of Opinion Scores

**SSIM** - Structural Similarity Index Measure

**STFT** - Short-Term Fourier Transform

**SVR** - Support Vector Regression

**TCD** - Trinity College Dublin

**VAD** - Voice Activity Detector

**ViSQOL** - Virtual Speech Quality Objective Listener

**ViSQOLAudio** - Virtual Speech Quality Objective Listener Audio

**VoIP** - Voice-over-IP

**W-PESQ** - Wideband Perceptual Evaluation of Speech Quality

**WSNR** - Weighted Signal-To-Noise Ratio

# 1 INTRODUCTION

The quality of speech and audio processing systems has been an important metric in the lives of most people in the developed world for many years. From the world's first transistor radio (the Regency TR-1) being developed in 1954, to the proliferation of modern-day portable music devices, it is clear that digital audio processing is the preference for audio consumers. In fact, it was found that 82% of university students in Canada used various digital portable audio devices in their daily lives. These included MP3 players, iPods, CD players and mobile phones (Ahmed et al., 2007). With regards to audio delivery, digital processing is clearly the dominant method whether it be mediums such as television, radio, cinema, how distortion is added to an artist's guitar or how a DJ will mix music at a live event. Music and sound effects are also seen as important immersive factors in video games (Sanders & Cairns, 2010). In some cases, immersion in the video game becomes so important that digital audio processing techniques are used to generate original music for the player on-the-fly (Epstein, 2016).

It is clear that speech and audio quality permeates quite a few facets of everyday life. Due to this, finding objective ways in which to measure this quality has been the subject of much research and investment. The research conducted in this paper aims to provide an answer to the following question: can machine learning techniques improve a currently existing objective speech quality metric?

## 1.1 Project Background

Digital speech processing techniques have been adopted worldwide. The most prominent example of this would be the adoption of the GSM (Global System for Mobile Communications) standard for cellular communications around the world. In the early 1980s, it was seen that analogue cellular telephone systems were being developed with incompatible standards across different states within the European Union. To remedy this, a standard was devised that would include criteria such as:

- A good subjective speech quality
- Compatibility with ISDN
- International roaming support

This would become the GSM standard that started commercial service in 1991 and went on to be adopted across the globe (Scourias, 1995). ISDN (Integrated Services Digital

Network) is a communications standard for the public switched telephone network which handles digital transmission of data, sound, and video.

For many years prior to GSM's introduction, speech signals have been transmitted across vast distances using the legacy telephone network service, which is now referred to as Plain Old Telephony Services (POTS). POTS provided a dedicated end-to-end connection for phone calls for decades. However, with the rise of digital technology and the need for digital networking, the POTS network connections (although never designed for digital signals) are becoming increasingly digitised. These are connections that consist of many hundreds of kilometres of twisted pair copper wire running from a service provider's hub to a customer's location and can provide a decent high-speed digital connection using technologies such as Asymmetrical Digital Subscriber Line (ADSL) (Kyees, McConnell, & Sistanizadeh, 1995). For markets where Internet access with even higher speeds are highly sought over, there has been a transition from POTS to high-quality IP telephony services. For example, ADSL growth has slowed down in Japan while growth in FTTH (Fibre-To-The-Home) has increased (Shinohara, 2005).

POTS has switched from an analogue telephone network to a digitised circuit-switched network over a period of many years. With the dawning of the 21st century, there has been a move from circuit-switched networks to packet-switched networks. These networks can carry voice, as well as data, over an Internet Protocol (IP) (Postel, 1981) network. With this has come the emergence of Voice-over-IP (VoIP) technologies (Goode, 2002). Specific architectures and protocols for VoIP have been developed by the International Telecommunication Union (ITU) (ITU-T, 1998b; Thom, 1996) and by the Internet Engineering Task Force (IETF) (Greene, Ramalho, & Rosen, 2000; Rosenberg et al., 2002). There are many platforms that utilise this technology, with Skype being the most popular. A study in the US in 2003 (the year Skype started operating) showed that some POTS networks were already suitable for high-quality VoIP. Although other POTS networks showed issues with reliability, network protocols, and router operation, actions could be taken to improve VoIP performance at the network and/or end-user systems (Markopoulou, Tobagi, & Karam, 2003). In October of 2012, after celebrating its 9th birthday, Skype recorded 45,469,977 concurrent users (Lunden, 2012). This clearly shows that the VoIP platform has continued to increase in popularity and provide an effective means of delivering speech signals across networks. Even with

the digitisation of telephony networks, customers still expect a high level of availability and quality (Kuhn, 1997).

The production, transmission, and perception of speech is referred to as the speech train. It can be simplified into 3 steps; These are generation, propagation, and reception. Speech can be generated by the vocal tract in humans and has varying characteristics such as loudness, frequency distribution, amplitude distribution, pitch rate, and syllabic rate. There can also be differences in how individuals speak. For example, their age, sex, mental state, and accent may all affect how the speech signal generated is perceived (Bedi et al., 2014; Goy, Kathleen Pichora-Fuller, & van Lieshout, 2016; Sharifzadeh, McLoughlin, & Russell, 2012). Upon receiving a speech signal, the listener must then be able to understand it. There are various reasons why it may not be perceived correctly due to the hearing ability and lexicon of the listener. The context of the speech signal may not be understood either.

As difficult as it can be for a human listener to perceive speech signals, one might wonder how does a computer system perceive speech if humans cannot always perceive it correctly? Due to the physical constraints of the human vocal system, such as breathing, there is an upper limit on how many words humans can transmit. As well as that, the rates of speech (or words per minute) can change based on factors such as age, sex, native, or second language spoken, speaking to strangers, or topics discussed (Yuan, Liberman, & Cieri, 2006). However, due to the nature of how speech is generated, speech signals are broken up into phones, phonemes, syllables, etc. This means that all speech, regardless of language or accent, has fundamental building blocks. This can be quite useful when getting computer systems to analyse speech signals in a digital format.

Two of the main factors to a listener of a speech signal are intelligibility and quality. A standard phone call will have a low quality signal but has quite a high rate of intelligibility. Voice transmission does not require high amounts of quality as human listeners can understand speech signals even in low quality environments. Intelligibility is a specific measurement. Generally, listeners will understand or not understand a speech signal. Intelligibility is necessary in the case of a phone call, but is not sufficient for quality. However, quality is related to intelligibility. If you have low intelligibility in a speech signal, the quality of that signal will also be low. It is a more subjective measurement than intelligibility. To give another example; a Skype call may have high

quality but drop packets at random intervals. A transmission such as this will drop the intelligibility of the speech signals transmitted. Thus, the perceived quality of the Skype call will also degrade.

There are various ways that one may measure intelligibility in a signal. Tests such as the diagnostic rhyme test (DRT), modified rhyme test (MRT), phonetically balanced word lists test, and ICAO spelling alphabet tests can all measure speech intelligibility (American National Standards Institute, 1989; House, Williams, Hecker, & Kryter, 1963; Martin, Champlin, & Perez, 2000; Schmidt-Nielsen, 1988). As discussed above, speech quality is a subjective measurement. This makes it more difficult to measure.

If a VoIP system wanted to constantly monitor the quality of the speech signals transmitted during a communication, it would not be practical to use a subjective test such as what is described above. Not only can the test take quite some time and be too costly, it is not possible for humans to react in time to inform the system that the speech signal quality had changed. However, there exist a number of objective sounds quality models that can give an objective value or score for a digitised speech signal in real-time. These models can be integrated into a VoIP system to provide immediate feedback of current speech quality.

## 1.2 Research Aims and Objectives

This research aims to evaluate the effectiveness of the use of machine learning techniques is in creating models that map the output values from objective sound quality models to a value that is understood to a user. With this in mind, and with the detailed review of existing literature explored below (Chapter 2), the effective Null Hypothesis is as follows: ***the accuracy of the ViSQOL objective speech quality metric cannot be improved using advancements taken from the newer ViSQOLAudio metric and training the output mapping function with a neural network and relevant speech data***. Both ViSQOL and ViSQOLAudio are objective sound quality models that will be explained in detail in Chapter 2.

## 1.3 Research Methods

The experiments conducted as part of this research will be empirical in nature. Primary data will be recorded from the experiments conducted. Some secondary data will be obtained from previous, relevant research for comparison purposes only. This research seeks to prove that accuracy can be improved in a selected objective sound quality model. It also aims to provide a basis for future work by comparing the machine learning techniques used, as well as the datasets used, to train the machine learning algorithms. The datasets used were obtained from external sources (International Telecommunication Union and Trinity College Dublin). The research is classified as empirical as it is aiming to increase accuracy which is a concrete and measurable metric.

## 1.4 Scope and Limitations

Within the scope of this study, two separate speech quality datasets were selected for testing. ViSQOL was originally trained with the ITU-T P Supplemental 23 database which contains audio samples from a number of different research laboratories (Hines, Skoglund, Kokaram, & Harte, 2015; ITU-T, 1998a). The database contains a collection of reference and degraded speech samples, as well as the MOS score that was given for the reference speech samples (ITU-T, 2006). It is designed to be used in the development of speech quality metrics. MOS scores are subjective quality measurements of sound samples. They are explained in greater detail in Chapter 2. The TCD-VoIP speech quality corpus was also selected as a candidate for training of machine learning models as it is presented in a simple format of WAV format speech files with associated MOS values for reference speech files (Harte, Gillen, & Hines, 2015). It is a valuable dataset as it was designed to emulate typical VoIP degradations.

While there are a relatively small number of available speech quality datasets that would be useful in the context of the research presented in this paper, the two datasets selected were deemed appropriate within the limitations of the experiments. The limitation in question being that the experiments need only be run on two separate datasets in order to prove results accurate. Additional datasets are not required to further prove whether improvements in accuracy occur or not.

The experiments were also limited to training one configuration of one type of machine learning model, as well as three separate configurations of a different machine learning model, for each dataset in order to test results on accuracy levels. Any further

experiments would be deemed unnecessary as the experiments conducted as part of this research should provide a basis of further work if required.

Full descriptions of the machine learning techniques and datasets used are available in Chapter 3, “Design / Methodology”.

## 1.5 Organisation of Dissertation

The remainder of this paper is organised into the following format:

- **Chapter 2 (“Literature Review”)** will explore previous research and provide a state-of-the-art analysis on how subjective sound quality is measured, the ITU’s role in providing standards, relevant objective sound quality models, detailed analysis of ViSQOL and the newer ViSQOLAudio, and machine learning techniques required in the scope of this research. Special attention is given to the ViSQOL model and the updated ViSQOLAudio objective sound quality models as the improvement of the accuracy of these models in the context of speech quality is the basis of this research. Similar research that uses machine learning techniques to predict sound quality is also discussed as it provides a precursor, or evidence of the relevance, for the experiments conducted as part of this paper.

- **Chapter 3 (“Design / Methodology”)** explores the datasets that will be used to train with machine learning techniques. These techniques are also detailed according to the manner in which they are utilised. Measurement of the accuracy of the models is also detailed in this section.

- **Chapter 4 (“Implementation / Results”)** details the results obtained from the experiments explained in Chapter 3. The results of each of the experiments run against both of the selected datasets are presented in a clear and concise manner. Tables and figures are presented as a visual guide to show the differences in accuracy for each of the machine learning techniques tested. The chapter concludes with a table presenting overall accuracy results for quick reference.

- **Chapter 5 (“Evaluation / Analysis”)** details the analysis of the results presented in Chapter 4. Observations on the resultant data are explored. Limitations on the research regarding datasets and machine learning techniques used are also explored.

- **Chapter 6 (“Conclusions and Future Work”)** summarises the entirety of the research presented in this paper, providing an explanation on how the research contributes to the body of research in objective sound quality metrics (with ViSQOL taking particular focus) and suggests further research that could be conducted from the results of this research.

# 2 LITERATURE REVIEW

This literature review will begin by focusing on research conducted on sound quality and objective sound quality models. Of these models, both ViSQOL and ViSQOLAudio will be discussed in the greatest detail as they are the main focus of this research. Finally, a review of published material on machine learning techniques used for analysing data for regression and classification (in the context of objective sound quality models) will be discussed.

While somewhat out of scope of the research presented in this paper, some research into the presence of machine learning in the mainstream media is provided in order to give context on the decreased level of expertise required for using machine learning and the rise in easily accessible machine learning tools. Thus, the experiments conducted as part of this research can be reproduced with a larger number of people.

Assumptions have been made that the reader is not versed in the somewhat specialised field of sound quality and objective sound quality models. Every effort has been made to explain the origins of the field as well as the workings of the objective sound quality models used as part of the experiments carried out in this research. It is, however, assumed that the reader has knowledge of machine learning techniques as there is little room to explain the roots of this field, nor the detailed mathematical formulae required, within the scope of this literature review. The machine learning techniques are discussed at a reasonable level of detail in the context of the research. Any readers interested in exploring the machine learning techniques used in greater detail will find vast amounts of reference material available from the usual sources.

**Note on the lexicon:**

Within the literature review presented, there is a mix of terminology used to describe features, metrics, and/or techniques. The “reference” and “degraded” signals discussed are generally deemed to be the inputs to a full reference objective sound quality model. The “reference” signal is an uncorrupted signal, while the “degraded” signal refers to a signal that has some kind of degradation or corruption. “Sound” quality generally refers to both speech and audio quality, while “speech” quality refers to only speech quality and “audio” quality refers to only audio (music) quality. “Similarity values” may refer to a specific metric such as the “NSIM” value (which is detailed later in this chapter) or any other similarity value outputted by various full reference objective sound quality

models. “MOS scores” or “MOS values” both refer to the same metric, which is a mean opinion score (which will be detailed later in this chapter). “Neural network” generally refers to an artificial neural network, rather than a naturally occurring biological neural network (e.g. in the human brain), in the context of this paper.

## 2.1 Measuring Sound Quality

Accurately measuring the quality of sound can be quite a complex task as it linked intrinsically to human hearing. While POTS networks traditionally provided a certain level of speech quality as standard, personal audio device manufacturers and codec creators constantly seek to improve and perfect sound recording and playback. Thus, there is no set standard for what is deemed a ‘perfect’ signal between different listeners. It is considered a subjective measurement.

### **2.1.1** Standardising Sound Quality Measurement

Since the 1990s, there have been great efforts to standardise sound quality measurement techniques. At present, the International Telecommunication Union (ITU) are seen as the governing body for speech and audio quality evaluation regulations with various regulations published.

In order to accurately measure subjective speech and audio quality, systematic subjective methodologies must be employed. In these test methodologies, a group of subjects (or listeners) are asked to rate the quality of a sample of audio that they are exposed to as part of the experiment. The most common type of subjective sound quality test is standardised in ITU-T P.800. Within the standard, an Absolute Category Rating (ACR) was devised to rate audio quality in subjective and objective testing using a Mean Opinion Score (MOS) (ITU-T, 1996). Each subject is asked to give their opinion of the quality of a sample of speech that has been played as part of the experiment. The subjects rate the stimulus played using a discrete scale shown in Table 2.1: MOS Scale with ‘1’ being the worst quality and ‘5’ being the best quality. In a typical test, signals within 5 seconds and 8 seconds are played to the subjects that contain two sentences, broken up by a 0.5 second period of silence, by the same speaker. 50 samples in this format are played to, and rated by, 24-32 subjects. The overall result of this test is the mean score of all the subject’s opinions on the speech quality, or the MOS.

| Score | ACR Listening Quality |
|---|---|
| 1 | Bad |
| 2 | Poor |
| 3 | Fair |
| 4 | Good |
| 5 | Excellent |

*Table 2.1: MOS Scale*

ITU-T P.800 also discusses alternative listening-opinion tests. These are the Degradation Category Rating (DCR) and the Comparative Category Rating (CCR) tests. With ACR, the order in which speech samples are presented to the listeners is not entirely important. However, with DCR, the order is relevant to the overall rating obtained. This is due to the fact that the degraded sample is presented immediately after the original sample and degradation is scored accordingly. The order at which samples are presented is random with CCR. Both DCR and CCR use separate discrete scales for measuring quality. DCR uses a scaling ranging from '1' (annoyingly degraded) to '5' (inaudibly degraded), while CCR uses a scaling ranging from '-3' (much worse) to '+3' (much better).

### 2.1.2 Mapping Quality Ratings to User Satisfaction Levels

Another quality scaling devised by the ITU, which will be discussed again further into this chapter, is the E-Model (ITU-T, 2003). The E-Model outputs a quality rating, *R*, which is a transmission rating factor measured between 0-100. Table 2.1 illustrates how user satisfaction can be mapped to a quality rating of *R*. Both Figure 2.1 and Table 2.2 illustrate how a MOS value can be mapped to an *R* transmission rating factor. Using this data, detailed MOS values can be assigned to user satisfaction ratings.

| R-value range | Speech transmission quality category | User satisfaction |
|---|---|---|
| $90 \leq R < 100$ | Best | Very satisfied |
| $80 \leq R < 90$ | High | Satisfied |
| $70 \leq R < 80$ | Medium | Some users dissatisfied |
| $60 \leq R < 70$ | Low | Many users dissatisfied |
| $50 \leq R < 60$ | Poor | Nearly all users dissatisfied |

*Table 2.2: Definition of categories of speech transmission quality (ITU-T, 1999, p. 2)*

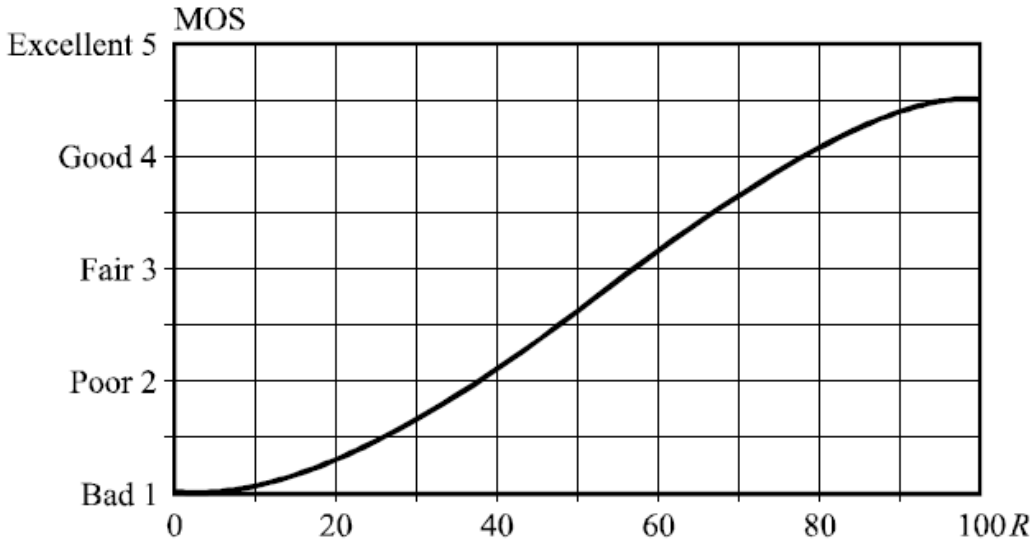



*Figure 2.1: MOS as a function of rating factor R (ITU-T, 2003, p. 16)*

| ***R*-value (lower limit)** | **$MOS_{CQE}$ (lower limit)** | **GoB (%) (lower limit)** | **PoW (%) (upper limit)** | **User satisfaction** |
|---|---|---|---|---|
| 90 | 4.34 | 97 | ~0 | Very satisfied |
| 80 | 4.03 | 89 | ~0 | Satisfied |
| 70 | 3.60 | 73 | 6 | Some users dissatisfied |
| 60 | 3.10 | 50 | 17 | Many users dissatisfied |
| 50 | 2.58 | 27 | 38 | Nearly all users dissatisfied |

*Table 2.3: Relation between R-value and user satisfaction (ITU-T, 2003, p. 16)*

A MOS score of above 4.3 corresponds to the best transmission quality for speech signals with high satisfaction levels from users. A MOS score of 4.0 to 4.3 corresponds to high quality with regular satisfaction levels. A MOS score of 3.6 to 4.0 corresponds to medium quality with some users dissatisfied. A MOS score of 3.1 to 3.6 corresponds to low quality with many users dissatisfied. A MOS score of 2.6 to 3.1 corresponds to poor quality with nearly all users dissatisfied. MOS ratings of 2.6 and below are not recommended. MOS ratings of above 4.0 match the levels of quality that are seen in POTS. These are the levels that VoIP must achieve in order to be a viable replacement for the levels of quality that are expected (e.g. with POTS). Using methods such as those described above to map user satisfaction to MOS values, Rämö was able to evaluate the subjective voice quality of various audio codecs (2010).

### 2.1.3 Advancements to MOS Quality Ratings

Fiedler et al. discusses in detail the relationship between Quality of Experience (QoE), which is perceived by users, and Quality of Service (QoS), which is affected by networks (2010). At times, the diverging views of some test subjects in QoE testing must be taken into account. Diversity of user ratings has been characterised and accounted for in previous research (Karapanos & Martens, 2007; Karapanos, Martens, & Hassenzahl, 2009). As a means of measuring subjective QoE, MOS experiments often do not focus on the lack of diversity in the subjective user ratings. Hoβfeld et al. argue that reporting just the MOS is not sufficient as even with large numbers of subjects, diversity of opinion occurs due to psychological factors such as expectation of quality levels, memory of quality previously experienced, and uncertainty on how to accurately grade a sample (2011). The authors proposed a Standard deviation of Opinion Scores (SOS) to be used in conjunction with MOS to rate the quality of the experiments undertaken to produce the MOS results. The SOS parameter a reflects rating diversity within an experiment. This parameter should be within a certain range in order to guarantee a level of quality in the experiment used to collect MOS values. Xu et al. investigated the QoE assumptions that are present in MOS calculations and found that the homogeneity that is required is lacking in practice (2011). The authors proposed a utility-based averaging for MOS called 'uMOS' which is designed to remove the 'unfairness' present in MOS averaging.

It is then clear that the subjective testing to acquire MOS values comes with some constraints such as sufficiently large number of test subjects to produce accurate results, equal control characteristics for each test subject, equal environmental conditions for each test, and a repeatable experimental procedure must be guaranteed (de Lima et al., 2008, p. 416). These issues can, and do, hinder the use of subjective measurement for speech and audio signals as they are costly and require much time to complete. Objective methods of acquiring quality measurements of speech and audio signals are therefore a more desirable option. Traditional automated methods to compute objective measurements for quality such as the use of Signal-To-Noise Ratio (SNR) have been shown to be lacking when it comes to accurately measuring speech quality (Hansen & Pellom, 1998). However, it must be noted that a weighted calculation of SNR (WSNR) has been shown to be effective at measuring the objective voice quality over cellular

networks (Karkhanechi, Gilhooly, & Soderstrand, 1998). Still more complex models are required to compute objective values for sound quality.

## 2.2 Objective Sound Quality Models

There are a number of different objective sound quality model types. The network channel and model estimates the quality of a signal without any reference to the input or output signal of a network (Moller et al., 2011). The no-reference signal-based models (e.g. P.563, ANIQUE+, LCQA) takes a reference from the output signal to estimate sound (ATIS, 2006; Grancharov, Zhao, Lindblom, & Kleijn, 2006; ITU-T, 2004). The parametric signal-based model (e.g. E-Model) monitors the state of the network itself to generate an estimated signal quality metric (ITU-T, 2003). The E-Model is a useful tool when analysing the perceptual quality of networks and had previously been used to evaluate to access networks in Pakistan (Mehmood, Jadoon, & Sheikh, 2005). Researchers have also proposed improvements to the model in the 'modified E-model' (Takahashi, Yoshino, & Kitawaki, 2004). There are additional full-reference signal-based models (e.g. PESQ, POLQA, ViSQOL, ViSQOLAudio) that reference both the input and output signals from the network in order to generate an objective speech signal quality score (Hines, Skoglund, et al., 2015; ITU-T, 2001b, 2011; Sloan, Harte, Kelly, Kokaram, & Hines, 2017). This research will focus on the full-reference signal-based models. In particular, it will focus on the ways in which objective speech quality measurement can be in improved in ViSQOL using advancements taken from ViSQOLAudio (an objective audio quality model) and the use of alternate mapping models.

PESQ (Perceptual Evaluation of Speech Quality) was standardised in ITU-T P.862. For many years, it has been the standard algorithm used for objective speech quality measurements. It involves three separate stages of processing: pre-processing, perceptual modelling, and cognitive modelling. The output of the PESQ model can be mapped to an objective MOS score with the following equation:

$$y = 0.999 + \frac{4.000}{1 + e^{Ax+B}} \quad (1)$$

$A = -1.4945$, $B = 4.6607$, while $x$ is the output from the PESQ model. $y$ is the objective MOS value.

PESQ can be used for signals within the range of 300 Hz to 3400 Hz. A wideband version of PESQ, (W-PESQ) was standardised in ITU-T P.862.2 and has a range of 50 Hz to 7000 kHz (ITU-T, 2007). With this model, the above equation is still used, but values for $A$ and $B$ are changed as follows: $A = -1.3669$ and $B = 3.8224$. Hu and Loizou showed in their research that, although PESQ performed best, both the LLR (log-likelihood ratio) and fwSNRseg (frequency-weighted segmental SNR) measures performed close enough to the PESQ results that they should be seriously considered as alternatives due to their significantly smaller computational cost (2008), (Tribolet, Noll, McDermott, & Crochiere, 1978). In recent years, POLQA (Perceptual Objective Listening Quality Assessment) was standardised in ITU-T P.863. It was specifically developed for HD Voice, 3G and 4G/LTE, and VoIP. It addresses some of the limitations of PESQ (warped speech and time alignment) while supplying objective quality values for narrowband, wideband, and super-wideband speech.

The ITU recommendation BS.1387 is a mixture of different individual contributions from various researchers (ITU-T, 2001a). It is referred to as PEAQ (Perceptual Evaluation of Audio Quality). It has two different versions: Basic and Advanced. It emulates some of the hearing properties of the human ear through software and produces an objective MOS value for quality. De Lima et al. argue that a neural network that is pre-trained and part of PEAQ may give unexpected results if presented with degraded samples that are not familiar to the network (2008). However, no evidence is brought forward to back up this argument. PEMO-Q is a psychoacoustic-based intrusive model that can evaluate both speech and audio across the audible spectrum that has shown better accuracy than that of PEAQ (Huber & Kollmeier, 2006).

## 2.3 ViSQOL

ViSQOL (Virtual Speech Quality Objective Listener) is a full-reference, signal-based, objective speech quality measurement model. It was brought about through the culmination of research carried out by the original authors (Hines, Počta, & Melvin, 2013; Hines, Skoglund, Kokaram, & Harte, 2012, 2013). It was designed to be deployable for a wide array of objective speech quality measurement situations, but is particularly suited to VoIP degradations in speech signals. The model works by examining time-frequency representations of the reference and degraded input signals in order to pinpoint particular VoIP degradations.

### **2.3.1** History of ViSQOL

The original inspiration for the ViSQOL model comes from research carried out on speech intelligibility by some of the original authors of the ViSQOL paper (Hines & Harte, 2010, 2012). In this work, auditory nerve discharge outputs were created by a model that simulated the workings of part of human ear (Zilany, Bruce, Nelson, & Carney, 2009). The model then outputted a neurogram which is essentially a visual representation of the time-frequency relationship of the neural firing activity. ViSQOL uses this research to create, not a neurogram, but a spectrogram of both the reference and degraded signals inputted into the ViSQOL model in order to contrast the differences between the signals. This is a somewhat unique approach to measuring the degradation of a signal, as most objective quality models will try to quantify how much distortion, noise, etc. is present in the degraded signal compared to the reference signal. A metric used for classifying the distance in similarity between the reference and degraded signal is used by ViSQOL dubbed the Neurogram Similarity Index Measure (NSIM). Speech intelligibility can be objectively measured by measuring the NSIM value between both the reference and degraded signals inputted into ViSQOL. NSIM is a variant of Structural Similarity Index Measure (SSIM) which is very widely used to measure image quality loss (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004). To put into simple terms, the input reference and degraded speech signals are essentially converted into an image that represents the spectrogram of the signals. These images are then compared to ascertain the difference between them in order to measure speech intelligibility. The reason spectrograms are used, rather than the neurograms from the research that inspired ViSQOL, is due to the computational cost of the model. A neurogram would increase the complexity of the model too much to be comparable with the computational cost of other objective speech quality metrics such as PESQ or POLKA.

### 2.3.2 The ViSQOL Model

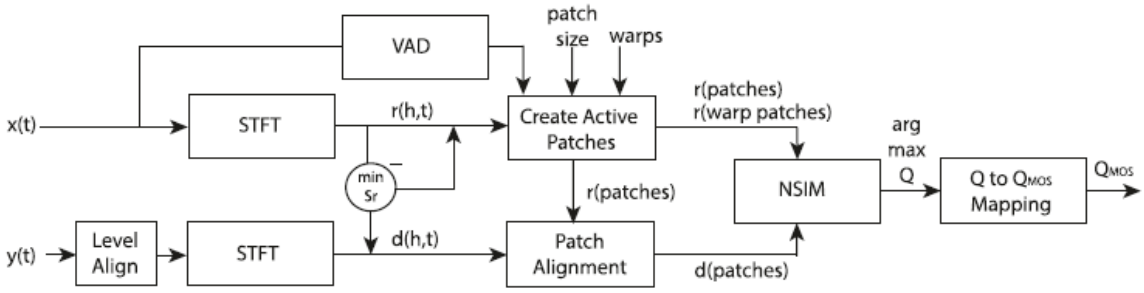



*Figure 2.2: High-level block diagram of ViSQOL (Hines, Skoglund, et al., 2015, p. 4)*

The ViSQOL model has five main parts; pre-processing, time alignment, predicting warp, similarity comparison, and mapping of similarity to objective quality. These are shown in Figure 2.2 above. In the first part of the model, pre-processing, the power of the degraded signal y(t) is scaled to match that of the reference signal x(t). A Short-Term Fourier Transform (STFT) spectrogram of both the reference and degraded signals is then created which are used as inputs to the next part of the model. These reference and degraded spectrograms are denoted r and d respectively.

The next part of the model involves time alignment of the reference and degraded signals. The reference signal is split into a number of different patches, each 30 frames in length. The signal is split into active patches by the use of a voice activity detector (VAD). The patches from the reference signal are then time aligned with the spectrogram from the degraded signal using NSIM to find the point at which the highest similarity occurs. These NSIM values (measured at the maximum of all the patches) are averaged across all the patches for the degraded signal to provide an overall NSIM value.

The ViSQOL model then proceeds to predict the warp on the signals. Different reference patches are created with warped values of 1% and 5% longer and shorter than the reference signal. The NSIM of each of the warped (and original) reference patches is measured against the degraded patch. The highest NSIM value is then used as the overall score for that patch. This is done because NSIM is better at picking up on time warped that the human ear is.

Next, the similarity comparison takes place. NSIM has been described above and is defined by the following equation (Hines, Skoglund, et al., 2015, p. 8):

$$Q(r,d) = l(r,d) \cdot s(r,d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_3}{\sigma_r \cdot \sigma_d + C_3} \quad (2)$$

Where $l$ is intensity, $s$ is structure, $\mu$ is the mean, $\sigma$ is standard deviation, $C_1 = 0.01L$, and $C_2 = C_3 = (0.03L)^2$. $L$ is defined as the intensity range of the reference input (in this case the reference signal). This equation is derived from workings of Wang et al. (2004). NSIM will return a result of '1' for a reference and degraded signal that are identical to each other, (i.e. no differences in the two input signals) and '0' for a reference and degraded signal that are completely dissimilar to each other (i.e. no part of the signals is similar). The mean of all the NSIM scores are calculated and returned as an overall NSIM score for the reference and degraded signal.

Finally, the ViSQOL model performs the mapping of the NSIM score to an equivalent objective MOS value. This mapping is done by a transfer function defined as follows (Hines, Skoglund, et al., 2015, p. 8):

$$clamp(Q_{MOS}, a, b) = \int \begin{matrix} m & if\ f(z) \leq m, \\ f(z) & a < f(z) \leq n, \\ n & if\ f(z) > n \end{matrix} \quad (3)$$

Where $Q_{MOS} = az^3 + bz^2 + cz + d$, $m = 1$, and $n = 5$. The coefficients are defined as thus: $a = 158.7$, $b = -373.6$, $c = 295.5$, and $d = -75.3$.

In the experiments conducted on ViSQOL by Hines et al., it was shown that it performed better than simpler metrics such as LLR and fwSegSNR and on par with both PESQ and POLKA (2015). It also noted that the mapping function had some issues for reference signals of lower quality with subjective MOS scores between '2' and '3'.

It is clear that the mapping from an NSIM score to an objective MOS value is likely the most important part of the ViSQOL model. Without proper mapping, the NSIM values calculated will be, in effect, wasted when using the model. While quite a lot of work has been conducted into distinguishing the differences between a reference and degraded signal, the mapping of ViSQOL does leave quite a lot to be desired.

## 2.4 ViSQOLAudio

ViSQOLAudio could be considered a 'newer' version of ViSQOL, even if it is specifically designed to rate the objective quality of music/audio samples rather than speech samples. The core of ViSQOLAudio contains some enhancements including a different approach to mapping NSIM scores to objective MOS values. The ViSQOL model described above was adapted to work with audio by removing the voice activity sensor, increasing the number of frequency bands evaluated, and removing the mapping

of NSIM values to MOS values in favour of a similarity scale between '0' and '1' (Hines, Gillen, et al., 2015). The reason the mapping was removed in this research was due to the fact that the mapping function used in ViSQOL was specifically designed for speech and would not give accurate objective MOS values for the new ViSQOLAudio model.

### 2.4.1 ViSQOLAudio Improvements

Sloan et al. built upon this work to further advance ViSQOLAudio (2017). There are five main improvements that were added to the ViSQOLAudio model with this research:

- Stereo audio samples are evaluated with both channels.
- Subframe misalignments of the input signals (reference and degraded) are compensated for.
- Fullband signals (audio) use a more appropriate filter bank.
- MOS values are outputted rather than a similarity score.
- MOS values are mapped from NSIM values with a machine learning model.

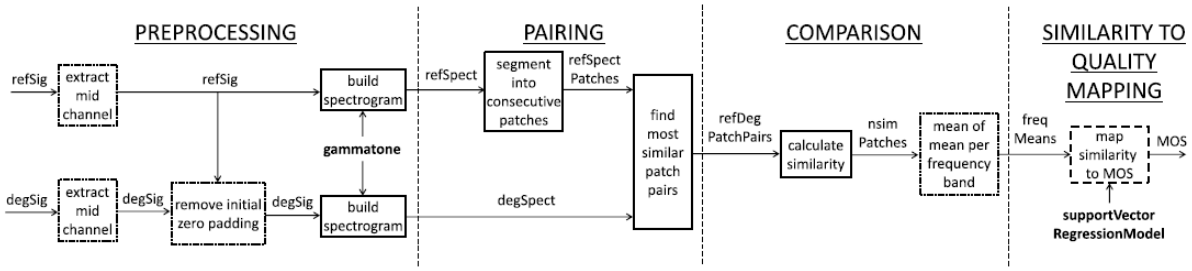



*Figure 2.3: High-level block diagram of ViSQOLAudio (Sloan et al., 2017, p. 3)*

Figure 2.3 shows the high-level diagram of the ViSQOLAudio objective sound quality model. It can be seen that this model is still quite similar to the original ViSQOL model shown in Figure 2.2. The pre-processing stage of the model now extracts the mid channel from the inputs. This is done in order to process information from both channels. The subframe misalignments are also compensated for in this stage. The pairing stage operates in a similar fashion to the original ViSQOL model. In the comparison stage, NSIM values are calculated across different frequency bands in order to evaluate similarity across a larger spectrum. The similarity to quality mapping stage uses a machine learning model to map NSIM values to an objective MOS value. Since the NSIM values are measured across the different frequency bands in this model, it allows for the machine learning model to look for matches for similarities throughout the

available frequency spectrum. The mapping model used is a support vector regression (SVR) model and allows for the mean of NSIM values of all similarity patches to be converted to an objective MOS value.

ViSQOLAudio generates an objective MOS value with the following process:

$$q = SVR\left(\frac{1}{M}\sum_{i=1}^{M}\Omega_i\right) \tag{4}$$

Where the objective MOS value is $q$, $M$ is the number of similarity patches, $\Omega$ is the NSIM values across the frequency bands, and $SVR$ is the support vector regression machine learning model.

In the experiments conducted on this version of ViSQOLAudio, it was shown that the model had superior results on two out of the three datasets used (Sloan et al., 2017). It came a very close second to a competing model in the third dataset. As an objective audio quality model, it is a viable alternative to PEAQ, POLQA and PEMO-Q.

| | | unmapped scaled to MOS-LQS | | | first order polynomial mapped | | | third order polynomial mapped | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dataset | model | R | OR | $\varepsilon$-RMSE | R | OR | $\varepsilon$-RMSE | R | OR | $\varepsilon$-RMSE |
| TCDAudio14 | ViSQOLAudio 2015 | 0.81 | **0.00** | 0.34 | 0.82 | **0.00** | 0.32 | 0.83 | **0.00** | 0.31 |
| TCDAudio14 | ViSQOLAudio | **0.93** | **0.00** | **0.18** | **0.93** | **0.00** | **0.17** | **0.93** | **0.00** | **0.17** |
| AACvOpus15 | ViSQOLAudio 2015 | 0.86 | **0.00** | 0.37 | 0.90 | **0.00** | 0.26 | 0.94 | **0.00** | 0.20 |
| AACvOpus15 | ViSQOLAudio | **0.94** | **0.00** | **0.19** | **0.94** | **0.00** | **0.18** | **0.96** | **0.00** | **0.14** |
| CoreSV14 | ViSQOLAudio 2015 | 0.79 | 0.13 | 1.16 | 0.79 | **0.00** | 0.54 | 0.80 | **0.00** | 0.54 |
| CoreSV14 | ViSQOLAudio | **0.89** | **0.05** | **0.99** | **0.91** | **0.00** | **0.34** | **0.91** | **0.00** | **0.29** |

*Table 2.4: Evaluation metrics for ViSQOLAudio (2015) Vs. ViSQOLAudio (2017) (Sloan et al., 2017, p. 10)*

Table 2.4 shows the improvements to linearity, accuracy and consistency between the original ViSQOLAudio and newer version of ViSQOLAudio presented by Sloan et al. (2017). The linearity, accuracy and consistency are calculated using Pearson's correlation coefficient (R), epsilon insensitive root means square error rate ($\varepsilon$-RMSE) and outlier ratio (OR) respectively (ITU-T, 2012). Therefore, although specifically tested with audio only, this iteration of ViSQOLAudio has improved upon its original form.

### **2.4.2** Objective Speech Quality to Objective Audio Quality

Objective sound quality models such as POLKA and ViSQOL have been adapted to work with audio. This is shown by the existence of the ViSQOLAudio and POLKA Music models. In the case of POLKA Music, the advancements made to create this model improved the objective speech metric of the model (Pocta & Beerends, 2015). This begs the question, could the improvements made to ViSQOLAudio resulted in

increased accuracy of the objective speech metric? Or could the SVR model be trained with speech data to improve the accuracy of the objective speech metric? Also, could a different machine learning technique be applied to the mapping model?

## 2.5 Machine Learning Techniques

The following section will detail the machine learning techniques that are already pre-existing, as well as those which will be used as part of the experiments conducted in this research, to create mapping functions from similarity values to objective sound quality values.

### **2.5.1** ViSQOLAudio and Support Vector Regression

ViSQOLAudio uses a support vector regression machine learning model in order to map NSIM values to objective MOS values. To achieve this, it uses LIBSVM and a training set of data to create the machine learning model (Chang & Lin, 2011). The model used is a ν-SVR with a radial kernel, $\nu = 0.6, \mathrm{cost} = 0.4$, and all other values set to defaults. There is precedent for the use of support vector regression machine learning models and sound quality. Liu et al. used SVR in order to predict diesel-engine related noise (2015). Additionally, Shen et al. used a support vector machine (which can be applied to classification and regression), a multiple linear regression model, and a neural network to predict vehicle interior sound quality (2010). Support vector machines have also been used with regards to audio steganography (Ozer, Avcibas, Sankur, & Memon, 2003).

### **2.5.2** Neural Networks Popularisation

Neural networks have become increasingly popular in academic and commercial use, and have been discussed widely in worldwide media in the last few years. As the technology has become more freely available and relatively easy to use, people have started to use it for a myriad of purposes. Its applications have ranged from the comical recipes that a badly trained neural network outputs (Alexander & Chambers, 2017), to commercial products generating original music to help customers add backing tracks to videos without having to worry about copyright issues on platforms such as YouTube (Chambers, 2017). Neural network training has even been streamed live on the popular video game streaming service ‘Twitch’. The blockbuster video game ‘Grand Theft Auto V’ has previously been discussed in the media as a training platform for self-driving cars (Matulef, 2017), but Harrison Kinsely, a Python programmer, has created a

convolutional neural network to teach a Python program how to drive within the virtual world of the video game while streaming its progress (O'Connor, 2017).

The most recent, high-profile example of a neural network making headlines would most likely be the Go match between South Korea's Lee Sedol (professional Go player with a rank of 9th dan) and the computer program 'AlphaGo' created by a subsidiary of Google called 'DeepMind' (Silver et al., 2016). The match, and the loss of one the world's greatest players one game to four, is seen as a milestone in the progress of AI (F.-Y. Wang et al., 2016). Many at the time compared it to a similar event almost 20 years' prior where a chess program, 'Deep Blue', developed by IBM, defeated the then World Chess Champion Garry Kasparov over six games in 1997 (Campbell, Hoane, & Hsu, 2002). However, the difference between the two programs is quite vast. Simply put, the reason 'Deep Blue' won the game was because computational power in computers had increased to a stage where the program could calculate and estimate every combination of moves in a game of chess. However, in a game of Go, there are vastly more combination of moves than any computer can calculate at present. This is why a neural network was developed and trained to learn from the data of games played by various professionals and from games it played with itself. The program has been shown to mimic human intuition in calculating its next move as it cannot possible know every move available. Therefore, it does not have the absolute 'right' move to play, just like a human player.

Thanks to these advancements, discussions, and popularisations of neural network techniques, experimentation with neural networks is far more accessible to the general public than it ever was. While a certain level of expertise in computing and mathematics is still required at present, that level is ever decreasing. Open source libraries, such as TensorFlow, are now available and can set up a neural network with GPU computational support in an extremely short amount of time compared to even a decade ago (Abadi et al., 2016).

### **2.5.3** ViSQOLAudio Accuracy Prediction

As discussed previously, ViSQOLAudio uses a machine learning model to predict the 'right' mapping between similarity values and objective MOS values. This research seeks to ask if a neural network model will improve accuracy over a support vector regression model. Narendra and Parthasarathy show in their research that neural

networks can be used effectively for identification (1990). Neural networks have already been used to predict sound quality in the automotive industry (H.-H. Lee & Lee, 2009; S.-K. Lee, 2008; S.-K. Lee, Kim, & Park, 2005; S.-K. Lee, Kim, & Lee, 2006; Y. S. Wang, Lee, Kim, & Xu, 2007). There is also a precedence to use neural networks with objective sound quality models as both PESQ and PEMO-Q use a neural network to map their respective model values to quality values (Huber & Kollmeier, 2006; ITU-T, 2001b).

## 2.6 Summary

The following section will detail a summary of the literature covered, any gaps discovered in the research as well as the research question for this paper.

### **2.6.1** Summary of Literature

While there are dozens of studies that discuss the workings of objective quality models and compare various models to one and other in terms of accuracy, it is clear that these objective quality models still have some way to go before they can be considered a true replacement for subjective sound quality measurement (Hines, Počta, et al., 2013; Hines et al., 2012; Hines, Skoglund, et al., 2013; Huber & Kollmeier, 2006; ITU-T, 2001b, 2011; Sloan et al., 2017). However, it is also clear that incremental advancements in accuracy are occurring with each of the newer models published over the last decade. An increasing amount of them are using machine learning techniques in order to map similarity values to objective sound quality values too (Hines, Gillen, et al., 2015; Huber & Kollmeier, 2006; ITU-T, 2001b). This may be a big factor in the recent increases in accuracy. From the research conducted on objective sound quality models that use machine learning techniques for mapping of similarity values to sound quality values, it is apparent that training with appropriate datasets has been a successful approach to the creation of such mapping models. Furthermore, with the popularisation of machine learning techniques (neural networks in particular), creating these type of mapping models has become much easier over the last few years (Abadi et al., 2016; Silver et al., 2016; F.-Y. Wang et al., 2016). That could be an additional factor into why there is a rise in the use of them being applied to objective sound quality models.

### **2.6.2** Gaps in Literature and Open Problems

While there appears to be some work done with regards to adding machine learning techniques into these full-reference objective sound quality models, there is quite a lack

of detail on the mapping models themselves. It appears that most researchers into objective sound quality models migrate image pattern recognition techniques into their models (Hines & Harte, 2010, 2012; Z. Wang et al., 2004). This leaves one questioning the level of expertise that the researches have with regards to the creation of machine learning models for mapping. However, as the research in this paper will show, this gap in knowledge may not actually be an issue as the level of expertise required for creating neural networks has lessened over time. Research from the perspective of a machine learning expert on the use of machine learning techniques in objective sound quality models would be most welcome. This could point out any issues or inconsistencies with the approach taken by most objective sound quality researchers. It could also lead to improvements in overall accuracy.

### **2.6.3** The Research Question

Through a thorough analysis of the available research into ViSQOL, its recent advancements, and similar full-reference objective sound quality models, the dominant motivation for the research presented in this paper was brought about: ***can the training of a new mapping model for ViSQOLAudio improve the accuracy levels (for speech quality) to levels greater than ViSQOL?***

# 3 DESIGN / METHODOLOGY

This section of the document will detail the methodology used to conduct the experiments required to answer the research question posed.

## 3.1 Method Used

In order to provide effective testing, datasets were selected that contain pairs of reference and degraded speech samples with accompanying Mean Opinion Scores (MOS). For the purposes of this research, such MOS values will be labelled as MOS-LQS (mean opinion score – listening quality subjective) as they are subjective values. The values outputted by the objective sound quality models tested will be referred to as MOS-LQO (mean opinion score – listening quality objective) as they are objectively calculated.

Accuracy and linearity are key metrics that were observed by this research in order to grade which objective sound quality model (or variant of) performs best (or has improved). To calculate accuracy, an F-score and root-mean-square error (RMSE) were calculated from the MOS-LQS and MOS-LQO results of each model tested. Linearity was measured using Pearson's correlation coefficient. This gives a measure of the linear relationship between both the objective and subjective values measured. As an added metric, the average difference in MOS score across the pairs of MOS-LQS and MOS-LQO results was generated to calculate the variance in results. These metrics allow for a measure of how often a model is accurate as well as how much it may deviate when it is inaccurate.

ITU-T P.1401 recommends that linearity, accuracy, and consistency be measured when evaluating objective sound quality models (ITU-T, 2012). However, the measurement of both accuracy and consistency will be used as part of this research. The only evaluation method taken from ITU-T P.1401 is the Pearson's correlation coefficient measurement. Further research could evaluate the results presented in Chapter 4 with the recommendations in ITU-T P.1401.

F-score (also referred to as F1-score or F-measure) is a measure of accuracy (Sokolova, Japkowicz, & Szpakowicz, 2006). Both the precision ($p$) and the recall ($r$) of the results (in the case of this research, the MOS-LQS and MOS-LQO values) are considered to compute the accuracy scoring. The sum of correct positive results divided by the sum of all positive results is $p$. The sum of the correct positive results divided by the sum of positive results that should have been returned is $r$. F-score is measured between '0' and

'1' with '0' being the worst accuracy rating and '1' being the best accuracy rating. It is the weighted average of *p* and *r* and is calculated with the following formula:

$$F_1 = 2 \cdot \frac{1}{\frac{1}{r} + \frac{1}{p}} = 2 \cdot \frac{p \cdot r}{p + r} \tag{5}$$

The root-mean-square error (RMSE) is used to calculate the true prediction error between both the subjective and objective MOS values recorded as part of this research. It is calculated with the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{n}} \tag{6}$$

Where $X_i$ is the predicted values (MOS-LQO) and $Y_i$ is the observed values (MOS-LQS). In the context of this research, the smallest value recorded will signify the smallest error in results between the datasets. RMSE can be viewed as a measure of how close the predicted data (MOS-LQO) is to the actual data (MOS-LQS) in the dataset.

Pearson's correlation coefficient ($R$) is a measure of the linear relationship between a selection of objective and subjective opinion scores (ITU-T, 2012). $R$ is calculated as thus:

$$R = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}} \tag{7}$$

Where $X_i$ is the MOS-LQS for speech sample $i$, $Y_i$ is the MOS-LQO for speech clip $i$, $\bar{X}$ is the mean MOS-LQS, $\bar{Y}$ is the mean MOS-LQO, and $N$ is the sum of speech samples in the dataset. With Pearson's, if a value is between '0' and '1', there is a positive correlation between the two sets of data. However, if a value is between '0' and '-1', then that shows a negative correlation between the two sets of data. In the context of this research, the ideal result between the two sets of data is as close to '1' as possible. Or in other words, the highest positive correlation between the two sets of data is ideal.

## 3.2 Sources of Data

Two relevant datasets were selected for use in the experiments conducted as part of this research. The first is the TCD-VoIP dataset (Harte et al., 2015). This dataset was chosen not only because it has not been used with ViSQOL or ViSQOLAudio in previous research, but due to the fact it represents data that is ideal for the development of objective sound quality models. Namely, samples of speech that have been degraded with various degradations that can be present in VoIP communications. For example, some of the degraded signals in the dataset contain echo, which can cause quality degradation in VoIP (Kostas et al., 1998, p. 20). The second dataset used was the ITU-T P Supplement 23 (P.Supp. 23) coded-speech dataset (ITU-T, 1998a). This dataset was developed for the 8 kbit/s codec (Recommendation G.729) designed by ITU-T and is a useful benchmark in speech quality measurement in objective VoIP speech quality models. Although relatively old at this point in time, it is still in use (M.-K. Lee & Kang, 2013). This dataset was also chosen as it has been used to train and benchmark ViSQOL in previous research (Hines, Skoglund, et al., 2015) and can effectively illustrate any differences in accuracy that this research may obtain.

## 3.3 Objective Sound Quality Models

As the purpose of this research is to improve on the accuracy of the ViSQOLAudio objective sound quality model with speech samples, both ViSQOL and ViSQOLAudio were chosen as the objective sound quality models to test accuracy levels against. ViSQOL was selected as an objective sound quality model in order to get a benchmark of MOS-LQO values for speech samples that need to be improved upon. ViSQOL was original designed for speech data and predicts quality relatively well compared to other objective sound quality models such as PESQ and POLQA. ViSQOLAudio is an updated version of ViSQOL with many improvements, but is specifically trained for audio/music samples. It was chosen as an objective sound quality model for this research as those improvements may lead to increased accuracy for speech samples as well.

The test data selected from within the chosen datasets were used as inputs both ViSQOL and ViSQOLAudio in order to gather accuracy results for comparison between the models.

## 3.4 Training ViSQOLAudio

ViSQOLAudio utilises a support vector regression machine learning model in order to map similarity scores taken from the reference and degraded audio samples inputted to a MOS-LQS value. Given the appropriate speech samples from both the TCD-VoIP and ITU-T coded speech datasets, ViSQOLAudio was then re-trained with a new support vector regression model using LIBSVM as well as three separate TensorFlow neural networks (Chang & Lin, 2011) (Abadi et al., 2016).

### **3.4.1** LIBSVM

When training a support vector regression model with LIBSVM, the subsets of data selected for training purposes from each of the dataset were run through the 'vanilla' ViSQOLAudio code with debug enabled. This allowed the capture of frequency band similarity scores for each of the pairs of reference and degraded speech samples. Each of these arrays of similarity scores were assigned an MOS-LQS value obtained from the dataset itself. The MOS-LQS value is our 'label' while each of the similarity scores measured are the 'values' associated with that 'label'. This allows for the creation of a support vector regression (SVR) model that can predict MOS-LQO values based on an input of an array of similarity scores. This approach was conducted as ViSQOLAudio used the exact same method to create a support vector regression model when training with audio data (Sloan et al., 2017).

### **3.4.2** TensorFlow

TensorFlow is an open source Python programming language library developed by Google that is utilised for the creation of machine learning algorithms. In the context of this research, it was used to create a deep neural network (DNN) to provide a model that can predict MOS-LQS values given an array of similarity scores. It is used as an alternative to the machine learning algorithm constructed using LIBSVM for ViSQOLAudio. As with the support vector regression model construction, MOS-LQS values obtained from each of the datasets were used as 'labels' and an associated array of similarity scores for each sample were used as 'values' when training the DNN. Three separate DNNs models were created with varying epoch training values.

## 3.5 Accuracy Measurement

The subsets of data from each of the datasets allocated for testing were used as inputs to the following objective sound quality models in order to obtain an F-score and RMSE value to measure accuracy between MOS-LQS and MOS-LQO values: ViSQOL, ViSQOLAudio, ViSQOLAudio with a new SVR model trained with speech samples, and ViSQOLAudio with three separate DNN models that replace the SVR model.

The results of each of the experiments was recorded in order to obtain F-score and RMSE accuracy readings for each model. These values were then used calculate the model that had the greatest accuracy in measuring the objective sound quality of speech samples with VoIP degradations.

## 3.6 Linearity and Average Difference in MOS-LQS and MOS-LQO Values

As an added measurement, the Pearson's correlation coefficient and average difference between the MOS-LQS and MOS-LQO for each of the models detailed was recorded. This allowed for a measurement of how much the results differed in some models. This is needed due to the fact that while accuracy may be relatively high in some cases, when the model calculates an incorrect MOS-LQO value, the difference in MOS-LQS and MOS-LQO may not be within tolerance levels of the user. Some models may give a lower accuracy but have incorrect MOS-LQO values within a user's tolerance level. Pearson's correlation coefficient can indicate a positive or negative correlation between the MOS-LQS and MOS-LQO values.

# 4 IMPLEMENTATION / RESULTS

The following section will detail and explain the results of the experiments that were conducted. It will be split into two sections. One detailing the experiments run against both ViSQOL and ViSQOLAudio with the TCD-VoIP dataset and another detailing the experiments run against ViSQOL and ViSQOLAudio with the ITU-T coded-speech dataset.

## 4.1 TCD-VoIP Dataset

The MOS-LQS scores that were provided with this dataset of speech samples were used to compute the accuracy of six different MOS-LQO scores. These six MOS-LQO scores were taken from the original ViSQOL MATLAB code, the original ViSQOLAudio, a version of ViSQOLAudio trained with a support vector regression model, as well as three versions of ViSQOLAudio replacing the support vector regression model with Tensorflow neural network models.

The dataset contains 5 different subset of files that are labelled ‘CHOP’, ‘CLIP’, ‘COMPSPKR’, ‘ECHO’, and ‘NOISE’. The models described in this section were trained with the first 4 subsets of the dataset (‘CHOP’, ‘CLIP’, ‘COMPSPKR’, and ‘ECHO’) and tested against the remaining subset of data (‘NOISE’). This subset of data used for testing contains 96 corresponding reference and degraded samples, while the subsets of data used for training contains 181 corresponding reference and degraded samples.

### **4.1.1** ViSQOL MOS-LQO Results

In order to obtain a reference of what this experiment hopes to improve on, the subset of data used for testing (‘NOISE’) was used as inputs for the original ViSQOL MATLAB code. This provided reference MOS-LQO scores. As the output of this code can return values between ‘0’ and ‘5’, with a precision of up to 9 decimal points, the MOS-LQO values outputted were rounded down to one decimal place values. Thus, a value of ‘4.922250734’ becomes ‘4.9’. This was done in order to ensure correct accuracy testing between the MOS-LQS (which have a precision of one decimal place) and MOS-LQO scores.

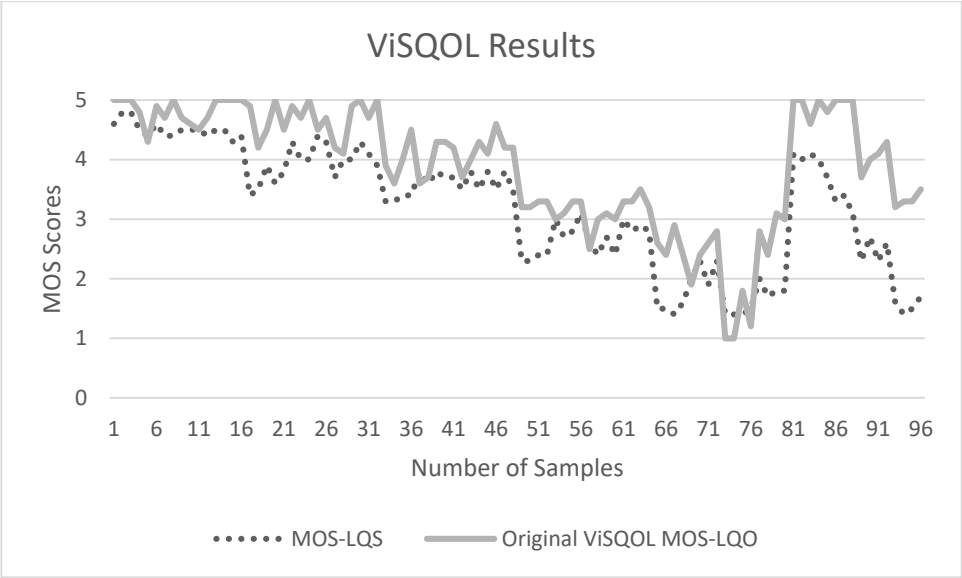



*Figure 4.1: TCD-VoIP Dataset ViSQOL Results*

Figure 4.1 illustrates a graphical representation of how closely the MOS-LQS values and ViSQOL MOS-LQO values align with each other. An F-score was calculated to distinguish how accurate the ViSQOL MOS-LQO scores are when compared to the MOS-LQS scores associated with the dataset. The result of this calculation was '0.021' (or 2.1%) accuracy. The RMSE was calculated as '0.86'. *R* was found to be '0.86'. The average difference in MOS scores between the two results was also calculated as '0.7'.

### **4.1.2** ViSQOLAudio MOS-LQO Results

The 'NOISE' reference and degraded speech samples were then used as inputs to the newer ViSQOLAudio in order to ascertain how a more updated version of the ViSQOL code (albeit trained specifically for music/audio) would respond to the same samples presented to the older ViSQOL code. ViSQOLAudio outputs MOS-LQO values ranging from '0' to '5' with a precision of up to 5 decimal places. In order to align correctly with the MOS-LQS results that are provided with the dataset, these results were rounded down to 1 decimal place of precision. Thus, a value of '4.94922' became '4.9'. This helped to calculate a more accurate F-score and RMSE between MOS-LQS provided by the dataset and MOS-LQO values that result from ViSQOLAudio.

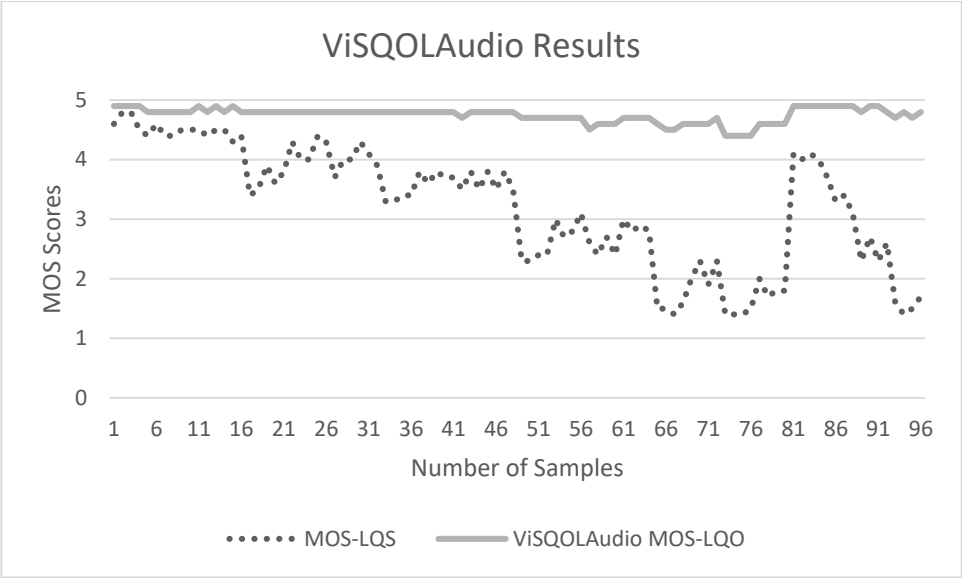



*Figure 4.2: TCD-VoIP Dataset ViSQOLAudio Results*

It can be immediately seen from Figure 4.2 that the accuracy is worse than that of ViSQOL as the MOS-LQO line does not cross the MOS-LQS line at any point. In fact, an F-score of '0' is calculated for the accuracy rating between the MOS-LQS and ViSQOLAudio MOS-LQO values. RMSE is '1.794'. $R$ is '0.759'. The average difference between the MOS-LQS and MOS-LQO results is calculated as '1.5'.

### 4.1.3 Training ViSQOLAudio for Speech using LibSVM

ViSQOLAudio was specifically trained with audio samples and is not designed to give accurate results for speech samples. It was trained using LIBSVM and all similarity scores that the code outputs are run through the trained model in order to output accurate results for audio samples. This section will detail how ViSQOLAudio was re-trained for speech samples with a new model.

Each of the remaining subsets of speech samples in the TCD-VoIP dataset (excluding the 'NOISE' subset as that is the test subset) were run through ViSQOLAudio in order to capture the similarity scores that are outputted by the code in debug mode. Using the MOS-LQS results from the dataset, a new support vector regression model was created using LIBSVM and the 'NOISE' test samples were run through this updated version of the code.

LIBSVM was given the following parameters when computing the model:

- The SVR is a nu-SVR with a radial kernel.

- The value of nu was set to ‘0.6’. This is the number of support vectors desired with respect to the number of samples in the dataset.
- The cost was set as ‘0.4’. This refers to the cost function of the model.
- All other parameters were kept as LIBSVM defaults.

These parameters were chosen as they are the same parameters used by Sloan et al. when training ViSQOLAudio for audio clips (2017).

### 4.1.4 ViSQOLAudio Trained with Speech Samples

With a new support vector regression model in place, the ‘NOISE’ subset of speech samples were run through the ViSQOLAudio code once more in order to see if there was any meaningful change to accuracy.

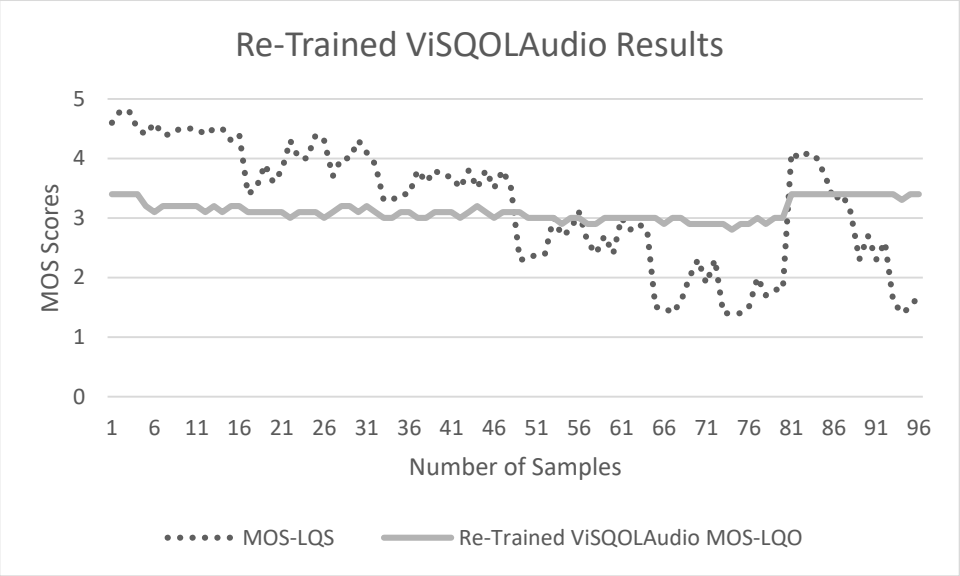



*Figure 4.3: TCD-VoIP Dataset Re-Trained ViSQOLAudio Results*

The data presented in Figure 4.3 achieves an F-Score of ‘0.031’, or 3.1%, and RMSE of ‘0.971’. As the last F-Score accuracy measurement was ‘0’, the re-training of the support vector regression model has increased the accuracy above that of the ViSQOL code (even if the RMSE has increased above the value recorded for that model). However, it can be seen that there is little variance in the results. There appears to be an averaging of the MOS-LQO results against the MOS-LQS scores. *R* was measured at ‘0.384’ and the average difference between each of the sets of MOS scores was ‘-0.1’.

### 4.1.5 Training ViSQOLAudio for Speech using Tensorflow

Using Tensorflow, a neural network was created to present a different model to that constructed by LIBSVM's support vector regression. The dataset was split into training and test data in the same manner that was used when re-training the support vector regression model. The 'NOISE' subset of speech samples were used as test data while all remaining speech samples were used to train the neural network. The data was split into 'labels' and 'features'. The 'labels' corresponded to the MOS-LQS values provided with the dataset. There were 51 unique 'labels' ranging from '0.0' to '5.0'. These represent the MOS-LQO values that will be outputted by the model. The 'features' were the similarity scores that were recorded when passing all reference and degraded speech samples present in the dataset through ViSQOLAudio. The neural network was tested against the 'features' (similarity scores) of the 'NOISE' subset of speech samples and provided 'labels' which are interpreted as MOS-LQO scores.

A neural network model was computed with the following parameters:

- Two hidden layers with 280 and 300 nodes respectively.
- Training rate of '0.01'.
- Three separate epoch training rates of 5,000, 50,000, and 1,000,000.

The neural network was trained with three separate training epoch values of 5,000, 50,000 and 1,000,000 in order to evaluate if accuracy increased as the cost function of the training reduced over time. These values themselves have no particular meaning other than to increase the time taken to train the model. Appendix A contains graphs illustrating the respective rise in epoch values and cost function reductions. While there is an initial drop in cost within the first few hundred epochs, it can clearly be seen that cost function eventually reduces to an acceptable level once enough training epochs are provided.

### 4.1.6 ViSQOLAudio with Tensorflow Neural Network Model

Once all three variants of the neural network with increasing training epoch times were completed, they outputted predicted MOS-LQO values for the 'NOISE' subset of reference and degraded speech samples. F-Score, RMSE, $R$, and average difference in MOS values were also calculated for each of the predicted array of MOS-LQO values.

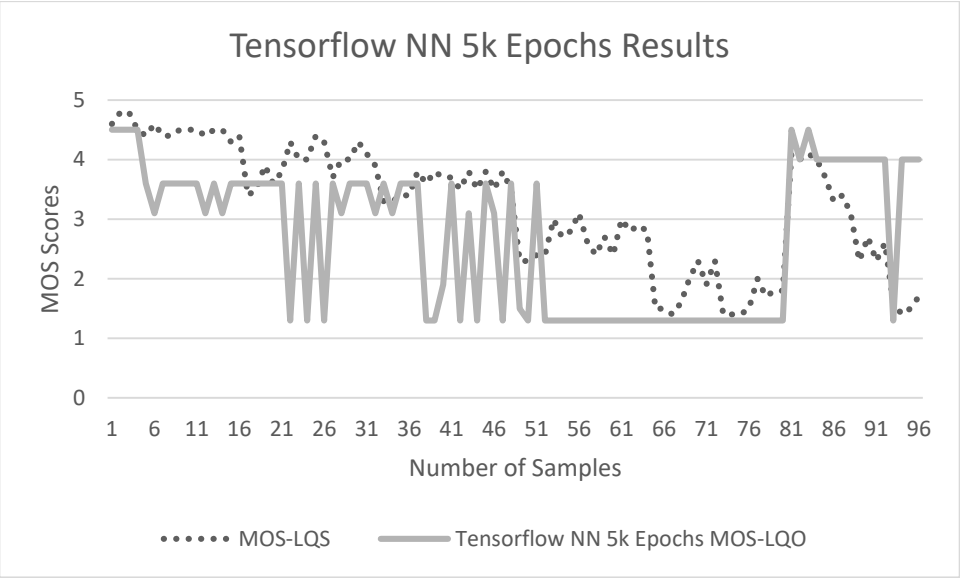



*Figure 4.4: TCD-VoIP Dataset Tensorflow NN 5k Epochs Results*

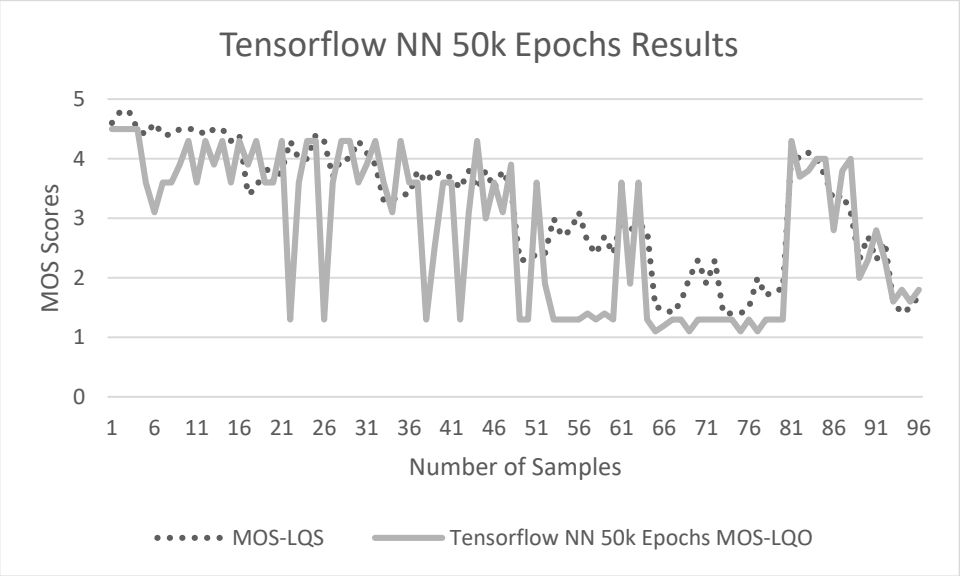



*Figure 4.5: TCD-VoIP Dataset Tensorflow NN 50k Epochs Results*

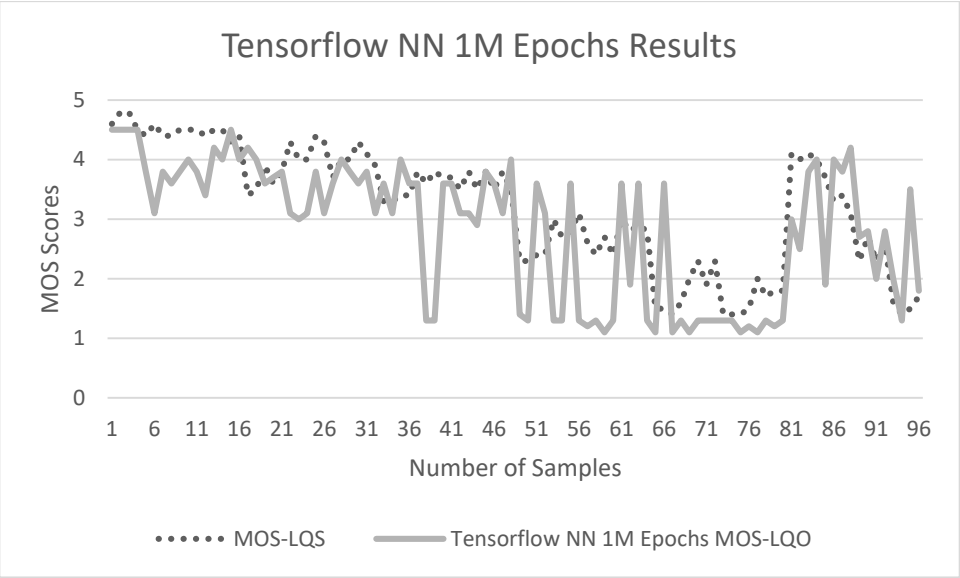



*Figure 4.6: TCD-VoIP Dataset Tensorflow NN 1M Epochs Results*

Figure 4.4, Figure 4.5, and Figure 4.6 illustrate the accuracy of the Tensorflow neural networks run with 5,000, 50,000, and 1,000,000 training epochs respectively. It can be seen that although the results are somewhat sporadic, they behave quite differently than those of the LIBSVM support vector regression models. The neural network models also achieved the greatest F-Score accuracy results of any of the models with '0.042' (or 4.2%), '0.042' (or 4.2%), and '0.052' (or 5.2%) respectively. RMSE was measured at '1.197', '0.872', and '0.877' respectively. *R* was measured at '0.551', '0.784', and '0.739' respectively. The average difference between MOS-LQS and MOS-LQO scores for each of the neural network models was '-0.5', '-0.4', and '-0.4' respectively.

### 4.1.7 TCD-VoIP F-Scores and Average Difference

This section illustrates the F-Scores, RMSE, $R$, and average difference between MOS-LQS and MOS-LQO results from each of the different models described above.

| **Models** | **F-Score (Accuracy)** | **RMSE (Accuracy)** | **Pearson's (Correlation)** | **Average Difference (in MOS)** |
|---|---|---|---|---|
| Original ViSQOL MOS-LQO | 0.021 | **0.856** | **0.861** | 0.7 |
| ViSQOLAudio MOS-LQO | 0 | 1.794 | 0.759 | 1.5 |
| New ViSQOLAudio MOS-LQO | 0.031 | 0.971 | 0.384 | **-0.1** |
| Tensorflow MOS-LQO 1 | 0.042 | 1.197 | 0.551 | -0.5 |
| Tensorflow MOS-LQO 2 | 0.042 | 0.872 | 0.784 | -0.4 |
| Tensorflow MOS-LQO 3 | **0.052** | 0.877 | 0.739 | -0.4 |

*Table 4.1: TCD-VoIP Dataset Results*

Table 4.1 shows the resultant F-Scores, RMSE, $R$, and average differences between MOS-LQS and MOS-LQO results for each model. The 'Tensorflow MOS-LQO 3' model (which was the neural network trained with 1,000,000 epochs) achieves the best accuracy with a score of '0.052'. The original ViSQOL model still achieves the lowest RMSE value and the highest $R$ (correlation) value. The least average difference between MOS-LQS and MOS-LQO results was the 'New ViSQOLAudio MOS-LQO' model with a difference of '-0.1'. This was the support vector regression model re-trained with the selected speech samples from the TCD-VoIP dataset.

## 4.2 ITU-T Coded Speech Dataset

As with the above section, this dataset was used to ascertain how accurate different models were at objectively measuring sound quality when compared to the subjective test measurements (or MOS-LQS values) supplied for each of the reference and degraded samples contained within the dataset. The ITU-T coded speech dataset contains data from three separate experiments on sound quality (‘EXP1’, ‘EXP2’, and ‘EXP3’). Of these three experiments, two contain reference and degraded samples of speech that have MOS-LQS values associated with them (‘EXP1’ and ‘EXP3’). The remaining dataset (‘EXP2’) contains reference and degraded speech samples but does not have associated subjective MOS values. Due to this, it was removed from this experiment, as associated subjective MOS values are a requirement with all speech samples used.

The original ViSQOL code trained a model using only one of the experiments contained in the dataset (‘EXP3’). For the purposes of this experiment, all remaining valid data contained within the dataset was used for training and testing of models generated (‘EXP1’ and ‘EXP3’). In order to make this a valid test against the original ViSQOL model, the same data that was used for testing that model was selected for this experiment (laboratory results labelled ‘O’ from experiment labelled ‘EXP3’). All remaining valid data was used for training of the models.

There were 1152 pairs of reference and degraded speech samples used for training of models and 203 used for testing trained models. In total, there were 1355 pairs of reference and degraded samples to utilise within this dataset.

The objective sound quality models that are used for testing against the subjective MOS values obtained from this dataset are as follows: ViSQOL, ViSQOLAudio, a re-trained ViSQOLAudio for speech samples, and three separate neural networks used as an alternative to ViSQOLAudio’s support vector regression model for similarity score to objective MOS value mapping.

### **4.2.1** ViSQOL MOS-LQO Results

The original ViSQOL objective sound quality model was used to test the reference and degraded speech samples from the selected test set within the dataset. However, it appears some issues with the processing of the samples occurred when conducting this experiment. While most pairs of reference and degraded speech samples were processed

without issue, a proportion of them simply outputted a MOS value of '1'. Although no error message was outputted by the code, it seems that there are some flaws to the MATLAB version of the model provided. This was unforeseen as this dataset has been previously documented to work with ViSQOL. It is unfortunate that these issues have been observed in the experiment, but it does not hinder the remaining results from other models.

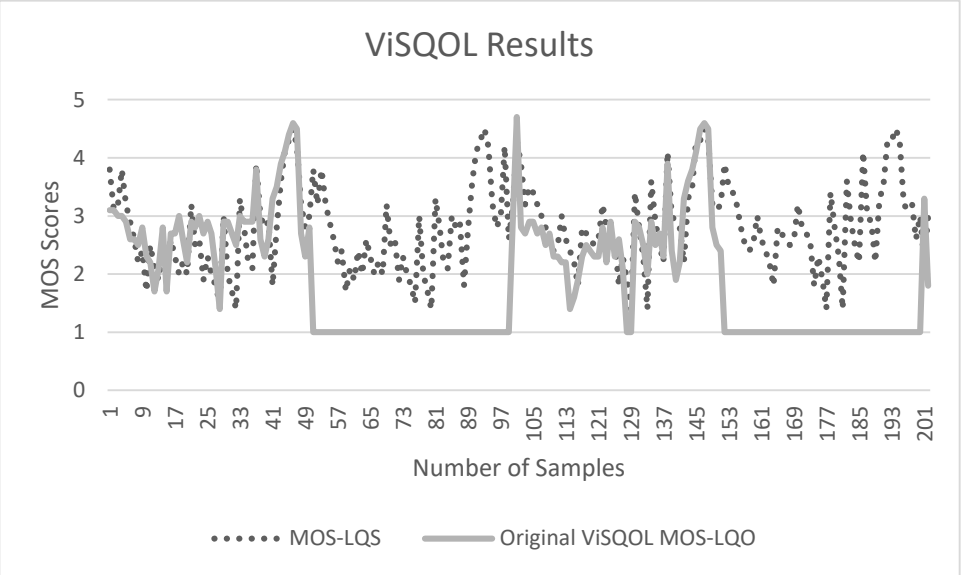



*Figure 4.7: ITU-T Dataset ViSQOL Results*

The issue described above is clearly visible in Figure 4.7. However, judging from the illustration of the data above alone, when the MATLAB model did work, it appeared to be quite accurate. From these results, an F-Score of '0.02' was observed. RMSE was measured at '1.398'. *R* was measured at '0.284'. Also, the average difference in MOS values between the two sets of values was '-0.9'. This value was more than likely dropped from what it should have been by the issues present with the model.

### 4.2.2 ViSQOLAudio MOS-LQO Results

ViSQOLAudio was tested with the ITU-T coded speech dataset in the same manner as the TCD-VoIP dataset was tested.

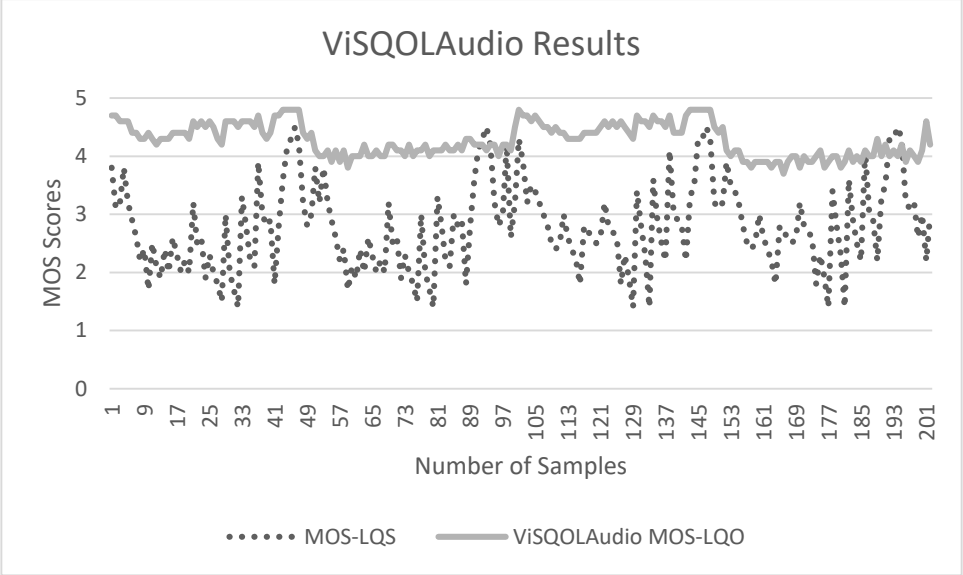



*Figure 4.8: ITU-T Dataset ViSQOLAudio Results*

Similar behaviour was observed with the MOS-LQO values outputted by ViSQOLAudio with both this dataset and the previous dataset. While there was some relative behaviours with the peaks and troughs between the MOS-LQS and MOS-LQO values, ViSQOLAudio was still outputting values far too high for these speech samples. The accuracy (F-Score) calculated from the MOS-LQS and MOS-LQO values was '0.015', or 1.5%. RMSE was measured at '1.664'. *R* was measured at '0.264'. The average difference between MOS-LQS and MOS-LQO values was '1.5'.

### 4.2.3 Training ViSQOLAudio for Speech using LIBSVM

The results from ViSQOLAudio were clearly not an improvement on the original ViSQOL when tested with this dataset (even with the errors seen with that model). Therefore, the support vector regression model used for mapping the similarity scores to MOS values was retrained with the 1152 pairs of reference and degraded speech samples designated for training in this dataset. This was done in the same manner that ViSQOLAudio was re-trained with the TCD-VoIP dataset described above.

Since the ITU-T coded speech dataset comes with MOS-LQS values with an accuracy of two decimal points, the support vector regression model was trained with these values. This is one difference to the support vector regression model trained with the TCD-VoIP

dataset which has MOS-LQS values with an accuracy of one decimal place. Due to the speed at which support vector regression models can be computed with LIBSVM, the more accurate values were not rounded down to one decimal place as it would affect accuracy.

### 4.2.4 ViSQOLAudio Trained with Speech Samples

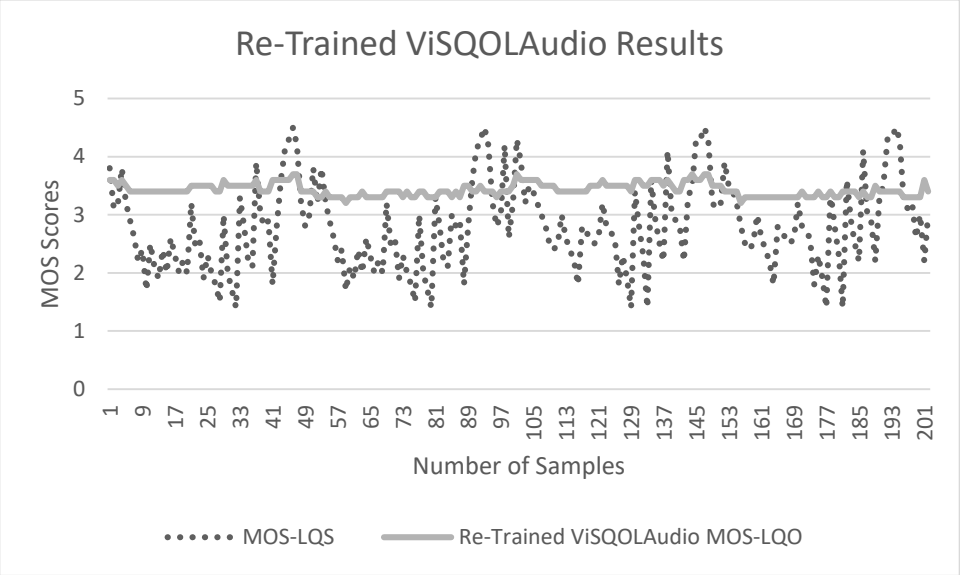



*Figure 4.9: ITU-T Dataset Re-Trained ViSQOLAudio Results*

The results from the re-training of the ViSQOLAudio support vector regression model show similar behaviour to that of the re-training conducted with the TCD-VoIP dataset. Namely that the resultant MOS-LQO values appear to be closer to an average of all MOS-LQS values than a proper mapping of MOS-LQS to MOS-LQO values with correct variance in results. The accuracy recorded for this model was '0.025' or 2.5%. RMSE was measured as '0.960'. *R* was measured as '0.377'. The average difference between MOS-LQS and MOS-LQO values was calculated at '0.6'.

### 4.2.5 Training ViSQOLAudio for Speech using Tensorflow

Since similar results were observed with this dataset to that of the previous dataset, namely poor mapping of similarity scores to MOS-LQO values with support vector regression methods, a neural network was constructed and trained as an alternative. The same Tensorflow neural network architecture was used to create a model for mapping MOS-LQS values supplied with the dataset to MOS-LQO values. The 'features' contained the similarity scores for each of the pairs of reference and degraded speech

samples designated for training in this dataset. The ‘labels’ contained the associated MOS-LQS values that were recorded for those paired speech samples.

It must be noted that due to the time it takes to train a neural network, the MOS-LQS values supplied with this dataset were rounded down from two decimal places to one decimal place. This meant that there were still 51 unique ‘labels’ in the neural network ranging from ‘0.0’ to ‘5.0’. Without this, there would be a need for 501 unique ‘labels’ ranging from ‘0.00’ to ‘5.00’ which would increase the time taken to run the neural network exponentially and was not feasible within the time limits of this research.

In a similar fashion to that of the neural network trained with the previous dataset (TCD-VoIP), it was observed that the cost function of the neural network sharply decreased initially within the first 100 training epochs (or iterations through the neural network). The cost function then steadily decreases to a more stable rate once the neural network is run for up to 1,000,000 training epochs. Three different training epoch values were assigned each of the three neural networks with the same values as that of the neural networks trained with the TCD-VoIP dataset. These were 5,000, 50,000, and 1,000,000 training epochs and are shown by graphs collected in Appendix B.

### 4.2.6 ViSQOLAudio with TensorFlow Neural Network Model

Each of the three neural networks models were then used as a replacement for the support vector regression model that maps similarity scores to MOS-LQO values. The three neural network models were run against the similarity scores from the reference and degraded speech samples in order to ascertain MOS-LQO values.

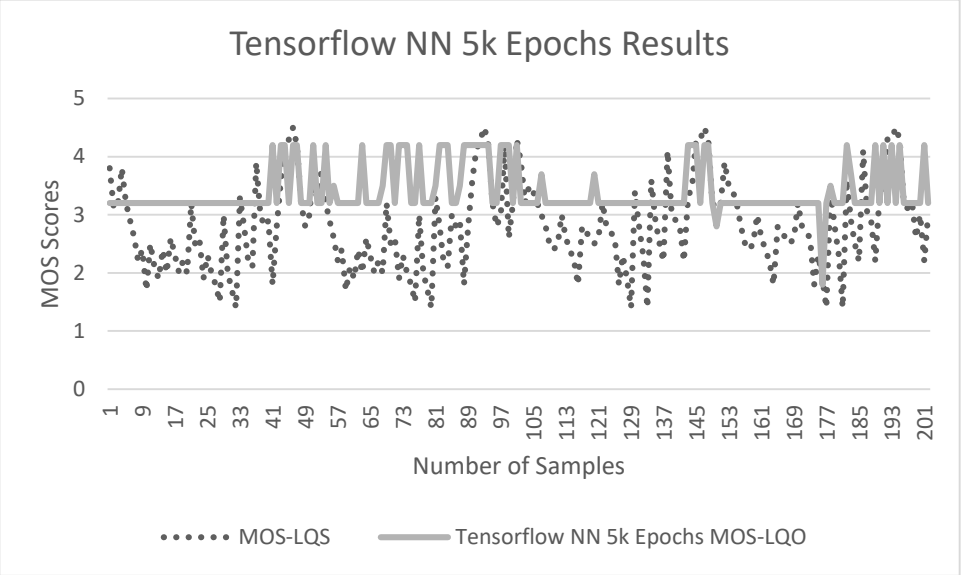



*Figure 4.10: ITU-T Dataset Tensorflow NN 5k Epochs Results*

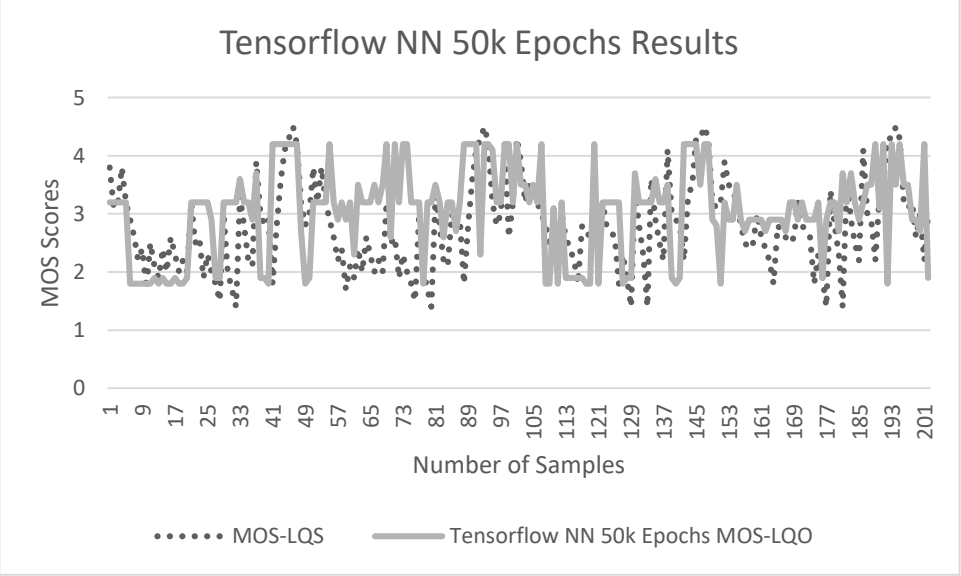



*Figure 4.11: ITU-T Dataset Tensorflow NN 50k Epochs Results*

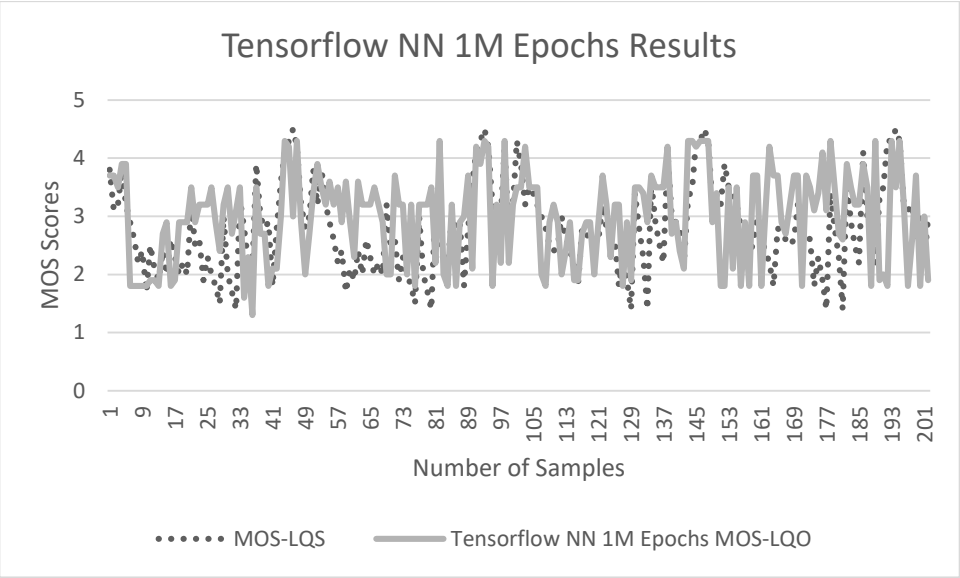



*Figure 4.12: ITU-T Dataset Tensorflow NN 1M Epochs Results*

Figure 4.10, Figure 4.11, and Figure 4.12 illustrate the results taken from each of the TensorFlow neural network models with epoch training values of 5,000, 50,000, and 1,000,000 respectively. The F-score accuracy measured for each of the neural network models was calculated as '0.094' (or 9.4%), '0.064' (or 6.4%), and '0.064' (or 6.4%) respectively. RMSE was measured as '0.947', '0.883', and '0.885' respectively. *R* was measured as '0.338', '0.382', and '0.372' respectively. Average difference between MOS-LQS and MOS-LQO results for each of the models was calculated as '0.6', '0.3', and '0.2' respectively.

### 4.2.7 TCD-VoIP F-Scores and Average Difference

This section illustrates the F-Scores and Average difference between MOS-LQS and MOS-LQO results from each of the different models described above.

| **Models** | **F-Score (Accuracy)** | **RMSE (Accuracy)** | **Pearson's (Correlation)** | **Average Difference (in MOS)** |
|---|---|---|---|---|
| Original ViSQOL MOS-LQO | 0.02 | 1.398 | 0.284 | -0.9 |
| ViSQOLAudio MOS-LQO | 0.015 | 1.664 | 0.264 | 1.5 |
| New ViSQOLAudio MOS-LQO | 0.025 | 0.960 | 0.377 | 0.6 |
| Tensorflow MOS-LQO 1 | **0.094** | 0.947 | 0.338 | 0.6 |
| Tensorflow MOS-LQO 2 | 0.064 | **0.883** | **0.382** | 0.3 |
| Tensorflow MOS-LQO 3 | 0.064 | 0.885 | 0.372 | **0.2** |

*Table 4.2: ITU-T Dataset Results*

From the overall F-score accuracy and average difference between MOS-LQS and MOS-LQO results shown in Table 4.2, it was seen that the highest F-score accuracy was seen for the TensorFlow neural network trained with 5,000 epochs. However, the average difference between the MOS-LQS and MOS-LQO values was relatively high (compared to other models) at '0.6'. The best values for RMSE and *R* were recorded for the TensorFlow neural network trained with 50,000 epochs. While the F-score accuracy measured for the remaining neural network models was recorded as '0.064' for each model, the lowest average difference between MOS-LQS and MOS-LQO results was seen with the neural network trained with 1,000,000 epochs with a value of '0.2'.

# 5 EVALUATION / ANALYSIS

The following section will discuss the evaluation, observations, strengths, and limitations of the results obtained in Chapter 4.

## 5.1 Evaluation of Results

This section of the paper will evaluate the results of each of the models (and variants thereof) that were tested with different datasets, as well as give an overall evaluation of the results captured from experiments.

### **5.1.1** Original ViSQOL Results

When testing the original ViSQOL model against the TCD-VoIP dataset, it was clear to see that it achieved reasonable accuracy. F-Score measures the rate at which the predicted MOS-LQO values were equal to the actual MOS-LQS values. Due the nature of objective sound quality models and the current amount of research into them, this value has not reached '1' or 100% accuracy for any current model. However, the aim of this research is to provide a stepping stone on the path to greater accuracy. F-Score is still a valuable metric to see how accurate a model has been though, especially in conjunction with the RMSE and Pearson's correlation value. In the case of ViSQOL tested with the TCD-VoIP dataset, F-Score accuracy was '0.021', quite a low accuracy level. Thus, one must look to the accompanying values for a fuller picture of the results. RMSE was '0.856' which shows that the model achieved the lowest rate of error when compared to the MOS-LQS values in the dataset. It is also seen that this model achieved the highest $R$ value, '0.861', which signifies the greatest positive linear relationship between the MOS-LQO and MOS-LQS values in the dataset. The average difference between MOS-LQO and MOS-LQS values was relatively low at '0.7'. This metric provides a fuller picture of the results in question. One could argue that RMSE gives a better idea of the difference, or error rate of the difference, between MOS-LQO and MOS-LQS values.

When testing the same model with the ITU-T P.Supp. 23 dataset, it is seen that results are not guaranteed in this experiment, as the MATLAB code provided seems to be flawed in some way. As mentioned in the previous chapter, the MATLAB code appears to output an MOS-LQO value of '1' when it experiences some unknown error. This is particularly frustrating as when the MATLAB code is working (as in the case with the experiment of the TCD-VoIP dataset), it works quite well. Also, this code has been

documented working with the ITU-T P.Supp. 23 dataset in the past (Hines et al., 2012). One can salvage some data from this failed experiment, however, as previous documentation in question does give an *R* value of '0.77' for testing the model against this dataset.

### **5.1.2** ViSQOLAudio Results

The results obtained from the ViSQOLAudio model were always predicted to be poor as the model itself is specifically trained for audio clips (Sloan et al., 2017). Hence the name – 'ViSQOLAudio'. However, the model itself does provide a substantial base on which to work upon. When tested with the TCD-VoIP dataset, some of the worst results for the overall experiment were recorded. The F-Score was '0' as no values of MOS-LQO were ever equal to the MOS-LQS values. RMSE was '1.794' which is quite a high error rate in a spectrum of values that ranges from '0' to '5'. It did however achieve a relatively high correlation rate with *R* measuring '0.759'. The average difference between MOS-LQO and MOS-LQS values was quite high at '1.5'.

When tested with the ITU-T P.Supp. 23 dataset, the F-Score was less than that of ViSQOL at '0.015'. The RMSE remained high at '1.664' and, surprisingly, the *R* value dropped considerably to '0.264'. While still a positive linear correlation, the large drop is surprising. The average difference between MOS-LQO and MOS-LQS values remained the same at '1.5'.

### **5.1.3** New ViSQOLAudio Results

As described in Chapter 4, the same test was run against the ViSQOLAudio model, once the SVR mapping model had been retrained with training data in both the TCD-VoIP and ITU-T P.Supp. 23 datasets.

In the case of the TCD-VoIP dataset, some of the measured metrics changed greatly. The F-Score accuracy level increased above that of even ViSQOL to '0.031', with a greatly decreased RMSE value of '0.971'. Surprisingly, the *R* value dropped to '0.384', which means that although there is less of a positive linear relationship between MOS-LQO and MOS-LQS, accuracy levels were still increased. The average difference between MOS-LQO and MOS-LQS values dropped significantly to '-0.1'. However, from Figure 4.3, it can be seen that the resultant SVR model has created an almost straight line prediction that essentially gives an average value of the test data. Thus, the significantly low average difference between MOS-LQO and MOS-LQO values.

When the same experiment was run with the ITU-T P.Supp. 23 dataset, similar behaviour was observed for the SVR mapping model when used with ViSQOLAudio. F-Score accuracy was slightly less than with the previous dataset at '0.025'. However, the RMSE was slightly lowered at '0.960' which shows less of a variance in results. The *R* value was roughly similar to that of the experiment on the previous dataset at '0.377'. The average difference between MOS-LQO and MOS-LQS values rose to '0.6'. Figure 4.9 shows that similar 'averaging' behaviour was observed with the predicted values from the mapping model.

### **5.1.4** Neural Network Results

As described in Chapter 4, three configurations of neural networks (with increasing epoch training rates) were trained with each dataset to produce mapping models for similarity values to MOS-LQO values.

With the TCD-VoIP dataset, each of the three models outputted higher F-Score accuracy measurements than that of ViSQOL. In fact, F-Score accuracy was at least doubled compared to ViSQOL in all cases. However, the model with just 5,000 training epochs outputted a large RMSE of '1.197', while the other two models outputted RMSE values quite close to that of ViSQOL which could mean more training was essential for the neural network. It is important to note that when the number of epochs increase, the risk of overfitting also increase. That is why there is not always a positive linear correlation between the number of epochs and accuracy levels. The model with just 5,000 training epochs outputted the lowest *R* value (between the three neural network models) of '0.551'. While still a substantial positive linear correlation between MOS-LQO and MOS-LQS values, both the models with 50,000 and 1,000,000 training epochs outputted *R* values of '0.784' and '0.739' respectively. These values are much closer to that of ViSQOL. Average difference between MOS-LQO and MOS-LQS values remained consistent across the three values with rising epoch training rates at '-0.5', '-0.4' and '-0.4' respectively.

### **5.1.5** Overall Evaluation of Results

With the TCD-VoIP dataset, the neural network models achieved the greatest levels of F-Score accuracy with comparable values of RMSE and *R* to that of ViSQOL. The average difference between MOS-LQO and MOS-LQS values was dropped slightly by the neural network models also. The SVR model trained with this dataset also achieved

a high level of accuracy, but still comes behind that of the neural network models. It also has increased levels of inaccuracy and difference in results measured by RMSE and *R* respectively.

Roughly the same observations were noted with the ITU-T P.Supp. 23 dataset. Some of the highest F-Score accuracy rates were measured with this dataset and the neural network mapping models. For the increasing levels of training epochs across the three neural network models, F-Score measurements of '0.094', '0.064', and '0.064' were recorded respectively. While the first neural network model had the greatest level of F-Score accuracy. It also had a higher error rate (RMSE), less correlation (*R*), and higher difference in results between MOS-LQO and MOS-LQS values than that of the other two models that were trained for longer. Essentially, it was 'right' in its predictions more of the time, but it came at a cost to error rate and correlation. Thus, the results for the neural network trained with 5,000 training epochs may be considered an outlier. The results from the neural networks trained with 50,000 and 1,000,000 training epochs are more consistent and still provide a good level of F-Score accuracy. The SVR model behaved roughly the same as the results given in the experiment with the previous dataset. F-Score accuracy increased with a re-trained SVR model, but RMSE error rate and *R* were still below expectations.

The correlation rate (*R*) given from previous research on ViSQOL still provides a higher positive rate of correlation than that of any of the new mapping models created as part of this research.

## 5.2 Observations from the Results

It is clear from the results presented in Chapter 4 that ViSQOL can and has been improved upon with new mapping models for ViSQOLAudio. The neural network models provided the greatest levels of F-Score accuracy, but they came at a cost to error rate (RMSE) and positive linear correlation between MOS-LQO and MOS-LQS results (*R*). Thus, the research question presented by this paper is answered: ***the training of a new mapping model for ViSQOLAudio improves accuracy levels (for speech quality) to levels greater than ViSQOL.*** While the error rate and positive correlation rates may be out of comfortable limits for some, this research shows that the improvements in ViSQOLAudio over ViSQOL can be used to improve upon ViSQOL.

Interestingly, it was found that the SVR models gave worse results than the neural network models. Although in a different field of research, DiPadua found that SVR models performed better than neural network models for pattern analysis (2016). Further research into the tuning of the selected machine learning models for this research may prove this to be true in the end.

## 5.3 Strengths of the Results

The results presented in this research show that machine learning techniques can, and are, valuable tools when attempting to increase the accuracy levels of objective sound quality models. The results can be used as a stepping stone for further research into this area as both the SVR and neural network models could possibly be tweaked to provide even greater accuracy results. Also, only two methods of machine learning techniques are explored as part of this research. There are many other techniques that could be applied to the objective sound quality models explored in this research for even greater results. Thus, the primary strength of the results presented in this research is the justification for feature selection as part of future work.

Accuracy data from the neural networks in this research is intended to be a first step in a rigorous investigation into the feasibility of neural network models for use with objective sound quality models. In particular, for use with ViSQOLAudio. However, the results do provide essential data on the successes and capabilities of how a neural network can handle the mapping of NSIM values to objective MOS values. The three different neural network models tested show that the data is not just a fluke, but instead show the relevance and capabilities of how a neural network model may be used in future with ViSQOLAudio. It also provides evidence that machine learning models can almost definitely be researched further for mapping functions in ViSQOLAudio.

The results themselves also show that the behaviour of the models are mostly similar across the two datasets used as part of the experiments undertaken in this research. It can be reasonably assumed that there are few outliers in the results presented. Even with a limited amount of datasets used, resultant data remains robust.

A secondary strength to the results presented as part of the experiments conducted in this research is that the models created can very easily be applied to currently existing objective sound quality models such as ViSQOLAudio. Once the mapping model has been created, mapping from similarity value to MOS-LQO takes very little time. Thus,

mapping models can be exchanged with ease for objective sound quality models such as ViSQOLAudio.

The final strength of the results is that the time taken to train models is relatively short. The process to train an SVR model takes quite a small amount of time and can be accomplished without specialised hardware. The neural network models with smaller training epochs can be trained and tested in a relatively short amount of time also. The SVM models created as part of this research took a number of seconds on an Intel Core i3-4030U CPU at 1.90 MHz. On the same hardware, the neural networks with 5,000 epochs took roughly five minutes to complete. Additional datasets could be merged to train and test additional models relatively easily for further advancements.

## 5.4 Limitations of the Results

The primary limitation of the results is that, as a measure of subjective QoE, the MOS values obtained as part of the datasets selected may in fact be flawed in the first place. Karapanos et al. argue against the common practice of averaging when analysing subjective measurements due to individuals' perceptions (2009). Karapanos and Martens suggest a different approach when modelling the difference in individuals' perceptions to avoid these issues (2007). With regards to MOS, it has also been argued that the measurement itself suffers from a lack of diversity in measuring subjective user rating. The SOS parameter has been proposed to be used in conjunction with MOS values to help avoid these issues (Hoβfeld et al., 2011). In practice, the required homogeneity for MOS has been found to be lacking and a utility-based averaging has been proposed to counteract this (Xu et al., 2011). With these issues present in the subjective MOS values that accompany the speech samples in the selected datasets, it is unclear whether or not the MOS-LQO values predicted by the models presented in this paper are trying to predict the 'right' value in the first place. However, within the scope of this research, it is assumed that the values for subjective MOS that are part of the datasets are correct. Further research could evaluate if these subjective values suffer from psychological factors or not. Further to this limitation, it has been shown that there are a number of biases that occur in quality listening tests (Zielinski, Rumsey, & Bech, 2008). For example, rating a sample 'OK' maps to various different levels of satisfaction across languages and cultures.

A secondary limitation to the results involves the allocation of training and test data from the datasets for the machine learning models. In both cases, the data used for training contained more samples that the test data, but there was no random selection of data. In the case of the TCD-VoIP dataset, four of the five subsets of data (that represented different VoIP degradations) were chosen for training data while the remainder were chosen for testing. It could be argued that if a random selection was used across all five of the subsets of data, better results may have been obtained. The results presented in this paper (for the TCD-VoIP dataset) show machine learning mapping models trained with various VoIP degradations, but then tested on a VoIP degradation which it was never trained on. While it performed well, it may not have been the best approach.

The ITU-T P.Supp. 23 dataset contains speech samples from different laboratories and experiments with a wide variety of degradations. The data that was selected for training and testing was safe from the concerns present for the TCD-VoIP dataset as they both contained similar degradations. However, a random selection for training and test data was not conducted as this research wanted to emulate the experiments carried out on ViSQOL (Hines et al., 2012). Further research could prove if a random selection of this dataset could provide better results.

A further limitation to the results presented in this paper involves the metrics used to evaluate the models. ITU-T P.1401 recommends that objective models should be assessed in terms of linearity, accuracy and consistency (ITU-T, 2012). It also recommends that first order and third order polynomial regressions are applied to the MOS-LQO data from the models. Hawkins formula should be used finding monotonically increasing polynomials for first and third order fits (Murray, Müller, & Turlach, 2016). Regression allows for a minimisation in RMSE and bias compensation for MOS-LQO data. Linearity should be measured with Pearson's correlation coefficient, accuracy should be measured with epsilon insensitive root mean square error (ε-RMSE) (which accounts for prediction error) and consistency should be measured with the outlier ratio (OR). The data collected as part of this research could be re-evaluated using the metrics presented as part of this recommendation in future work.

Another important limitation of the results involves the tuning of the SVR and neural network models. One study concluded that "SVR probably has greatest use when the dimensionality of the input space and the order of the approximation creates a

dimensionality of a feature space representation much larger than that of the number of examples" (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997, p. 160). In the context of this research, it is worth investigating the dimensionality of the data presented to the SVR model as an input for training. Similarity values related to subjective MOS values may not have enough dimensionality for the SVR model chosen. Further tweaking of the SVR model may be required. Chromagram, mel-scaled spectrogram, mel-frequency cepstral coefficients, spectral centroid tonal centroid features, and/or zero crossing rate data of the input signals could add dimensionality to the models. However, this would also come at a cost of computational processing of the objective sound quality model in question.

Hornik et al. concluded in their research that "standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy, in a very specific and satisfying sense. We have thus established that such "mapping" networks are universal approximators" (1989). From this it can be seen that if a neural network mapping model does not achieve the desired amount of success, it must be due to inadequate learning rates, lack of hidden layer nodes, or that the relationship between input and output is not deterministic enough.

For the neural network created in the experiments detailed in this research, two hidden layers were used. Funahashi's research showed that any mapping can be achieved using just one hidden layer, while for pattern recognition, two hidden layers are used (1989). Similar research (based on Kolmogorov's theorem) to prove how the use of two hidden layers in a neural network has the capability to provide universal approximation has also been conducted (Hornik, 1991; Kůrková, 1992). Further research could prove if different configurations or numbers of hidden layers may improve the accuracy of the neural network mapping models created using neural networks.

The final limitation to the results is this: ViSQOLAudio was created to process, analyse, and compare audio samples, not speech samples. In the process of converting ViSQOL to ViSQOLAudio, elements important to speech detection, such as the voice activity detector, were removed (Hines, Gillen, et al., 2015). It was also trained to work with audio samples only. While this research has successfully re-trained the model to work with speech samples, it is not known how the presence of original elements from ViSQOL (such as the voice activity sensor) would have had an effect on the results.

Further research could investigate how the addition of such elements effect ViSQOLAudio's performance with speech data.

# 6 CONCLUSIONS

This section of the paper details the conclusions found as part of the research. The overview of the research, a definition of the research problem, the design, experimentation, evaluation and results is given. The contributions and impact are discussed, and recommendations for future research are given.

## 6.1 Research Overview

A thorough analysis of the available research on the measurement of subjective sound quality, various objective sound quality models (ViSQOL and ViSQOLAudio in particular), and machine learning techniques (support vector regression and neural networks) was conducted. The research implies that, while subjective sound quality testing may have some psychological flaws in its methodology, objective sound quality testing has progressed greatly over the last decade. However, accuracy levels when measuring objective quality of sound still require some improvement.

## 6.2 Problem Definition

From the research conducted, it was evident that many useful advancements have occurred in the ViSQOL objective sound quality model as it has advanced from an objective speech quality model (ViSQOL) to an objective audio quality model (ViSQOLAudio). These advancements include the calculation of a more comprehensive similarity score between reference and degraded input signals, and the addition of a support vector regression machine learning model to map similarity scores to objective MOS values. Thus, the research problem presented by this paper asked the following: ***can the training of a new mapping model for ViSQOLAudio improve the accuracy levels (for speech quality) to levels greater than ViSQOL?***

## 6.3 Design/Experimentation, Evaluation and Results

Two separate speech datasets (TCD-VoIP and ITU-T P.Supp. 23) were selected for both the training and testing of machine learning algorithms, as well as testing of standard ViSQOL and ViSQOLAudio models for benchmark results. The support vector regression model used by ViSQOLAudio (to map similarity scores to objective MOS values) was retrained in two separate experiments using the two selected datasets. The support vector regression model was also replaced by a neural network model, with three different configurations, and tested with the same datasets. Thus, results were obtained

for ViSQOL, ViSQOLAudio, ViSQOLAudio with a re-trained SVR mapping model, and ViSQOLAudio with three different neural network mapping models.

The metrics used for evaluating models were as follows; F-Score and RMSE for accuracy, Pearson's correlation coefficient for linearity, and average difference between real and predicted values as a simpler linearity/accuracy measure.

The results of the experiments showed that ViSQOLAudio, in some cases, can be used to give objective speech quality metrics with greater accuracy than ViSQOL. Both the re-trained support vector regression model and the neural network models showed greater accuracy than ViSQOL with the TCD-VoIP dataset. However, this came at a slight cost to linearity and error rate. There were issues with using ViSQOL on the ITU-T P.Supp. 23 dataset, but some results from previous research on ViSQOL were obtained for comparison. They showed that linearity values for the new machine learning models did not reach that of ViSQOL's original results.

## 6.4 Contributions and Impact

This research sought to evaluate if the accuracy levels of objective sound quality models could be increased with the use of machine learning techniques. The general conclusion brought forward from this research is that both support vector regression and neural network models are superior to a standard transfer function when mapping similarity scores to objective MOS values. It was also seen that the ViSQOLAudio model can be altered, with relative ease, to accommodate for both speech and audio quality measurement.

The viability of machine learning mapping models for ViSQOL and ViSQOLAudio is proven by this research and provides a ground basis for future improvements to the accuracy of the objective sound quality models.

## 6.5 Future Work and Recommendations

Two speech quality datasets were chosen to be used as part of this research (Harte et al., 2015; ITU-T, 1998a). However, there exists additional appropriate datasets that could be used as part of further research, such as the NOIZEUS narrowband noisy speech corpus (Hu & Loizou, 2007). Future research should make every attempt to collect as much viable data as reasonably possible and combine the datasets to create much larger and varied training and test data. The experiments presented in this research could be repeated with the increased training dataset size in order to evaluate if training time

decreases and/or accuracy improves. Shalev-Shwartz and Srebro argue that, in the case of support vector machines, the increase of training set size should decrease the time taken to train the model as well as decrease the error rate (2008).

As part of this research, a support vector regression model was trained with both of the selected speech datasets. The parameters used to train this machine learning model were purposefully identical to the parameters chosen when ViSQOLAudio was trained with audio datasets (Sloan et al., 2017). It has been shown that if there is not enough dimensionality to the training data, results of the SVR model may be poor (Drucker et al., 1997). Further research into the SVR mapping models used for ViSQOLAudio should explore different approaches to how the model itself is trained. Different configurations should be tested in order to ascertain the optimum parameters to train a mapping model with the available datasets. The addition of extra dimensionality to the training and test data should be explored to ascertain if an increase in accuracy is observed with the resultant SVR mapping model.

Three separate neural networks were trained and tested with the two chosen datasets in this research. The only difference in their configuration was the amount of training epochs allocated to the models. This resulted in different training times for each of the models. It has been proven that neural networks with two hidden layers are ideal for pattern recognition systems (Funahashi, 1989; Hornik, 1991; Kůrková, 1992). Therefore, optimisations to the configuration of the neural networks used should be explored as part of further research. This should help to improve upon the accuracy results obtained in this research.

Lastly, due to the relative ease of replacing the mapping model used for ViSQOLAudio, it is recommended that the authors of ViSQOLAudio seriously consider the implementation of a separate model for objective speech quality measurement based on this research. A ‘ViSQOLSpeech’ model based on the core mechanics of ViSQOLAudio, that re-introduces the voice activity detector, and uses a mapping model based on this research could potentially be a very accurate measure of objective speech quality.

# 7 Bibliography


Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … others. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv Preprint arXiv:1603.04467*. Retrieved from https://arxiv.org/abs/1603.04467

Ahmed, S., Fallah, S., Garrido, B., Gross, A., King, M., Morrish, T., … Pichora-Fuller, K. (2007). Use of portable audio devices by university students. *Canadian Acoustics*, *35*(1), 35–52.

Alexander, L., & Chambers, I. (2017, April 21). What would an AI make you for dinner? – tech podcast. *The Guardian*. Retrieved from https://www.theguardian.com/technology/audio/2017/apr/21/what-would-an-ai-make-you-for-dinner-tech-podcast

American National Standards Institute. (1989). Method for measuring the intelligibility of speech over communication systems (ANS S3.2-1989). *New York: American Standards Association*.

ATIS, A. (2006). 0100005-2006: auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality,". *American National Standards Institute*.

Bedi, G., Cecchi, G. A., Slezak, D. F., Carrillo, F., Sigman, M., & De Wit, H. (2014). A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, *39*(10), 2340–2348.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, *134*(1), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Chambers, L. A. I. (2017, April 14). Can a neural network compose music you want to hear? – Tech podcast. *The Guardian*. Retrieved from https://www.theguardian.com/technology/audio/2017/apr/14/can-a-neural-network-compose-music-you-want-to-hear-tech-podcast

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

de Lima, A. A., Freeland, F. P., de Jesus, R. A., Bispo, B. C., Biscainho, L. W., Netto, S. L., … others. (2008). On the quality assessment of sound signals. In *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on* (pp. 416–

419). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/4541443/

DiPadua, J. (2016). Support Vector Machines and Artificial Neural Networks: Assessing the Validity of Using Technical Features for Security Forecasting. Retrieved from http://arrow.dit.ie/scschcomdis/93/

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161). Retrieved from http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf

Epstein, M. (2016, August 9). How 'No Man's Sky' composes completely original music for every player. Retrieved 6 May 2017, from https://www.digitaltrends.com/gaming/no-mans-sky-music/

Fiedler, M., Hossfeld, T., & Tran-Gia, P. (2010). A generic quantitative relationship between quality of experience and quality of service. *IEEE Network, 24*(2). Retrieved from http://ieeexplore.ieee.org/abstract/document/5430142/

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*(3), 183–192. https://doi.org/10.1016/0893-6080(89)90003-8

Goode, B. (2002). Voice over Internet protocol (VoIP). *Proceedings of the IEEE*, *90*(9), 1495–1517. https://doi.org/10.1109/JPROC.2002.802005

Goy, H., Kathleen Pichora-Fuller, M., & van Lieshout, P. (2016). Effects of age on speech and voice quality ratings a. *The Journal of the Acoustical Society of America*, *139*(4), 1648–1659.

Grancharov, V., Zhao, D. Y., Lindblom, J., & Kleijn, W. B. (2006). Low-complexity, nonintrusive speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1948–1956.

Greene, N., Ramalho, M., & Rosen, B. (2000). *Media Gateway control protocol architecture and requirements*. Retrieved from https://www.rfc-editor.org/rfc/pdfrfc/rfc2805.txt.pdf

Hansen, J. H., & Pellom, B. L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *ICSLP* (Vol. 7, pp. 2819–2822). Retrieved from https://pdfs.semanticscholar.org/4974/18c70971c8d990e2edf989d6f05675b7c23a.pdf

Harte, N., Gillen, E., & Hines, A. (2015). TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on* (pp. 1–6). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/7148100/

Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., & Harte, N. (2015). ViSQOLAudio: An objective audio quality metric for low bitrate codecs. *The Journal of the Acoustical Society of America*, *137*(6), EL449-EL455.

Hines, A., & Harte, N. (2010). Speech intelligibility from image processing. *Speech Communication*, *52*(9), 736–752. https://doi.org/10.1016/j.specom.2010.04.006

Hines, A., & Harte, N. (2012). Speech intelligibility prediction using a Neurogram Similarity Index Measure. *Speech Communication*, *54*(2), 306–320. https://doi.org/10.1016/j.specom.2011.09.004

Hines, A., Počta, P., & Melvin, H. (2013). Detailed comparative analysis of PESQ and VISQOL behaviour in the context of playout delay adjustments introduced by VOIP jitter buffer algorithms. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on* (pp. 18–23). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6603195/

Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N. (2015). ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1). https://doi.org/10.1186/s13636-015-0054-9

Hines, A., Skoglund, J., Kokaram, A., & Harte, N. (2012). ViSQOL: The virtual speech quality objective listener. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on* (pp. 1–4). VDE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6309421/

Hines, A., Skoglund, J., Kokaram, A., & Harte, N. (2013). Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 3697–3701). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6638348/

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

House, A. S., Williams, C., Hecker, M. H., & Kryter, K. D. (1963). Psychoacoustic speech tests: A modified rhyme test. *The Journal of the Acoustical Society of America*, *35*(11), 1899–1899.

Hoβfeld, T., Schatz, R., & Egger, S. (2011). SOS: The MOS is not enough! In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 131–136). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6065690/

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*(7), 588–601.

Hu, Y., & Loizou, P. C. (2008). Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238. https://doi.org/10.1109/TASL.2007.911054

Huber, R., & Kollmeier, B. (2006). PEMO-Q: A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(6), 1902–1911. https://doi.org/10.1109/TASL.2006.883259

ITU-T, R. (1996). Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva, Switzerland*, 800.

ITU-T, R. (1998a). ITU-T Coded-Speech Database. *International Telecommunication Union, Geneva, Switzerland*.

ITU-T, R. (1998b). Packet-based multimedia communication systems. *International Telecommunication Union, Geneva, Switzerland*, *323*.

ITU-T, R. (1999). 109, Definition of Categories of Speech Transmission Quality. *International Telecommunication Union, Geneva, Switzerland*.

ITU-T, R. (2001a). Methods for objective measurements of perceived audio quality. *International Telecommunication Union, Geneva, Switzerland*, 1387–1.

ITU-T, R. (2001b). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *International Telecommunication Union, Geneva, Switzerland*. Retrieved from ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG/MIRROR/ITS/vqeg.its.bldrdoc.gov/%252E%252E/%252E%2

52E/Documents/VQEG_Ghent_Sept08/MeetingFiles/T-REC-P.910-199909-I!!MSW-E.doc

ITU-T, R. (2003). Recommendation G. 107 The E-model, a computational model for use in transmission planning. *International Telecommunication Union, Geneva, Switzerland*.

ITU-T, R. (2004). P. 563: Single-ended method for objective speech quality assessment in narrow-band telephony applications. *International Telecommunication Union, Geneva, Switzerland*.

ITU-T, R. (2006). *P. 800.1, Mean opinion score (MOS) terminology*. International Telecommunication Union, Geneva, Switzerland.

ITU-T, R. (2007). P. 862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, Geneva, Switzerland*.

ITU-T, R. (2011). Perceptual objective listening quality assessment. *Message Sequence Charts (MSC96)*.

ITU-T, R. (2012). ITU-T Rec. P.1401: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *International Telecommunication Union, Geneva, Switzerland*.

Karapanos, E., & Martens, J.-B. (2007). Characterizing the diversity in users' perceptions. In *IFIP Conference on Human-Computer Interaction* (pp. 515–518). Springer. Retrieved from http://link.springer.com/10.1007/978-3-540-74796-3_50

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2009). Accounting for diversity in subjective judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 639–648). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1518801

Karkhanechi, H. M., Gilhooly, D., & Soderstrand, M. A. (1998). Objective measurement of voice quality for cellular mobile phones. In *Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on* (pp. 161–165). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/759460/

Kostas, T. J., Borella, M. S., Sidhu, I., Schuster, G. M., Grabiec, J., & Mahler, J. (1998). Real-time voice over packet-switched networks. *IEEE Network, 12*(1), 18–27.

Kuhn, D. R. (1997). Sources of failure in the public switched telephone network. *Computer*, *30*(4), 31–36.

Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, *5*(3), 501–506. https://doi.org/10.1016/0893-6080(92)90012-8

Kyees, P. J., McConnell, R. C., & Sistanizadeh, K. (1995). ADSL: a new twisted-pair access to the information highway. *IEEE Communications Magazine*, *33*(4), 52–60.

Lee, H.-H., & Lee, S.-K. (2009). Objective evaluation of interior noise booming in a passenger car based on sound metrics and artificial neural networks. *Applied Ergonomics*, *40*(5), 860–869. https://doi.org/10.1016/j.apergo.2008.11.006

Lee, M.-K., & Kang, H.-G. (2013). Speech quality estimation of voice over internet protocol codec using a packet loss impairment model. *The Journal of the Acoustical Society of America*, *134*(5), EL438-EL444.

Lee, S.-K. (2008). Objective evaluation of interior sound quality in passenger cars during acceleration. *Journal of Sound and Vibration*, *310*(1–2), 149–168. https://doi.org/10.1016/j.jsv.2007.07.073

Lee, S.-K., Kim, B.-S., & Park, D.-C. (2005). Objective evaluation of the rumbling sound in passenger cars based on an artificial neural network. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *219*(4), 457–469. https://doi.org/10.1243/095440705X11112

Lee, S.-K., Kim, T.-G., & Lee, U. (2006). Sound quality evaluation based on artificial neural network. *Advances in Natural Computation*, 545–554.

Liu, H., Zhang, J., Guo, P., Bi, F., Yu, H., & Ni, G. (2015). Sound quality prediction for engine-radiated noise. *Mechanical Systems and Signal Processing*, *56–57*, 277–287. https://doi.org/10.1016/j.ymssp.2014.10.005

Lunden, I. (2012, October 14). Skype Reaches A 45M Concurrent User Peak, And What Looks Like A New Stage Of Momentum. Retrieved 7 May 2017, from http://social.techcrunch.com/2012/10/14/skype-reaches-a-45m-concurrent-user-peak-and-what-looks-like-a-new-stage-of-momentum/

Markopoulou, A. P., Tobagi, F. A., & Karam, M. J. (2003). Assessing the quality of voice communications over internet backbones. *IEEE/ACM Transactions on Networking*, *11*(5), 747–760. https://doi.org/10.1109/TNET.2003.818179

Martin, F. N., Champlin, C. A., & Perez, D. D. (2000). The question of phonetic balance in word recognition testing. *JOURNAL-AMERICAN ACADEMY OF AUDIOLOGY*, *11*(9), 509–513.

Matulef, J. (2017, April 17). Grand Theft Auto 5 is being used to teach driverless cars. Retrieved 28 May 2017, from http://www.eurogamer.net/articles/2017-04-17-grand-theft-auto-5-is-being-used-to-teach-driverless-cars

Mehmood, M. A., Jadoon, T. M., & Sheikh, N. M. (2005). Assessment of VoIP quality over access networks. In *Internet, 2005. The First IEEE and IFIP International Conference in Central Asia on* (p. 5–pp). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/1598195/

Moller, S., Chan, W.-Y., Cote, N., Falk, T., Raake, A., & Waltermann, M. (2011). Speech Quality Estimation: Models and Trends. *IEEE Signal Processing Magazine*, *28*(6), 18–28. https://doi.org/10.1109/MSP.2011.942469

Murray, K., Müller, S., & Turlach, B. (2016). Fast and flexible methods for monotone polynomial fitting. *Journal of Statistical Computation and Simulation*, *86*(15), 2946–2966.

Narendra, K. S., & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, *1*(1), 4–27.

O'Connor, A. (2017, May 18). Watch this AI learn to drive inside Grand Theft Auto V. Retrieved 28 May 2017, from https://www.rockpapershotgun.com/2017/05/18/grand-theft-auto-v-livestream-ai-learns-to-drive/

Ozer, H., Avcibas, I., Sankur, B., & Memon, N. D. (2003). Steganalysis of audio based on audio quality metrics. In E. J. Delp III & P. W. Wong (Eds.) (pp. 55–66). https://doi.org/10.1117/12.477313

Pocta, P., & Beerends, J. G. (2015). Subjective and Objective Assessment of Perceived Audio Quality of Current Digital Audio Broadcasting Systems and Web-Casting Applications. *IEEE Transactions on Broadcasting*, *61*(3), 407–415. https://doi.org/10.1109/TBC.2015.2424373

Postel, J. (1981). Internet protocol.

Rämö, A. (2010). Voice quality evaluation of various codecs. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (pp.

4662–4665). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/5495201/

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., … Schooler, E. (2002). Request for Comments 3621—SIP: Session Initiation Protocol, Jun. 2002. *Internet Engineering Task Force*, 8–14.

Sanders, T., & Cairns, P. (2010). Time perception, immersion and music in videogames. In *Proceedings of the 24th BCS interaction specialist group conference* (pp. 160–167). British Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=2146327

Schmidt-Nielsen, A. (1988). Evaluating the intelligibility of different speech degradations using the ICAO spelling alphabet. *The Journal of the Acoustical Society of America*, *84*(S1), S15–S15.

Scourias, J. (1995). Overview of the global system for mobile communications. *University of Waterloo*, *4*. Retrieved from http://www.di.unisa.it/~ads/corso-security/www/CORSO-9900/a5/gsmreport/gsmreport.pdf

Shalev-Shwartz, S., & Srebro, N. (2008). SVM optimization: inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning* (pp. 928–935). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1390273

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2012). A Comprehensive Vowel Space for Whispered Speech. *Journal of Voice*, *26*(2), e49–e56. https://doi.org/10.1016/j.jvoice.2010.12.002

Shen, X.-M., Zuo, S.-G., Li, L., & Zhang, S.-W. (2010). Interior sound quality forecast for vehicles based on support vector machine. *Zhendong Yu Chongji(Journal of Vibration and Shock)*, *29*(6), 66–68.

Shinohara, H. (2005). Broadband access in Japan: Rapidly growing FTTH market. *IEEE Communications Magazine*, *43*(9), 72–78.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Sloan, C., Harte, N., Kelly, D., Kokaram, A. C., & Hines, A. (2017). Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio. *IEEE Transactions on Broadcasting*, 1–13. https://doi.org/10.1109/TBC.2017.2704421

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australian conference on artificial intelligence* (Vol. 4304, pp. 1015–1021). Retrieved from https://vvvvw.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-006.pdf

Takahashi, A., Yoshino, H., & Kitawaki, N. (2004). Perceptual QoS assessment technologies for VoIP. *IEEE Communications Magazine*, *42*(7), 28–34.

Thom, G. A. (1996). H. 323: the multimedia communications standard for local area networks. *IEEE Communications Magazine*, *34*(12), 52–56.

Tribolet, J. M., Noll, P., McDermott, B., & Crochiere, R. (1978). A study of complexity and quality of speech waveform coders. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78.* (Vol. 3, pp. 586–590). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/1170567/

Wang, F.-Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., … Yang, L. (2016). Where does AlphaGo go: from Church-Turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, *3*(2), 113–120.

Wang, Y. S., Lee, C.-M., Kim, D.-G., & Xu, Y. (2007). Sound-quality prediction for nonstationary vehicle interior noise based on wavelet pre-processing neural network model. *Journal of Sound and Vibration*, *299*(4–5), 933–947. https://doi.org/10.1016/j.jsv.2006.07.034

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

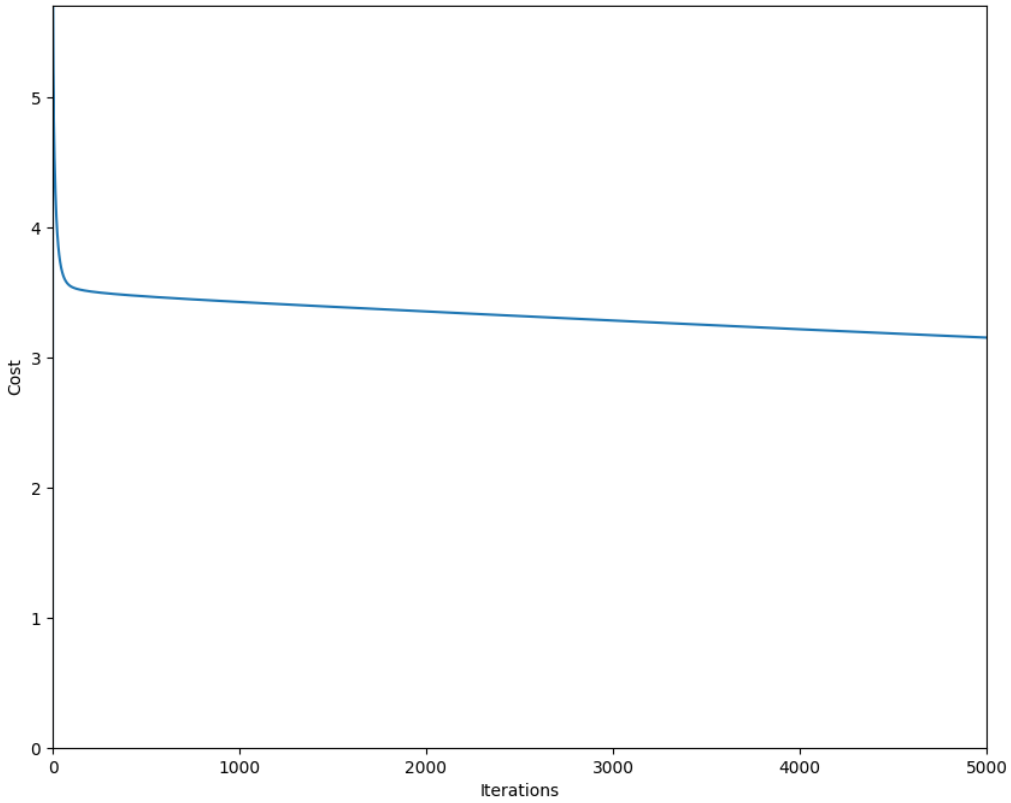
Xu, J., Xing, L., Perkis, A., & Jiang, Y. (2011). On the Properties of Mean Opinion Scores for Quality of Experience Management (pp. 500–505). IEEE. https://doi.org/10.1109/ISM.2011.88

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *INTERSPEECH*. Retrieved from https://pdfs.semanticscholar.org/f7b1/8ae6558b10305e504a24fcd79749aad93d3a.pdf

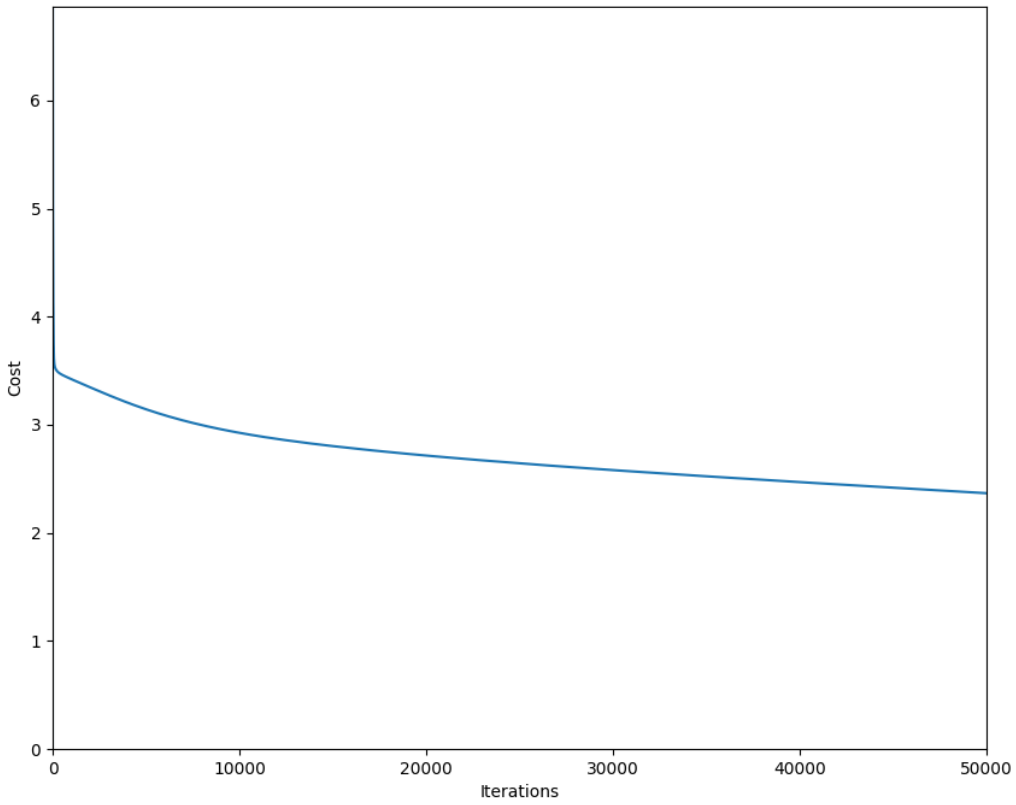
Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. *Journal of the Audio Engineering Society, 56*(6), 427–451.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., & Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America, 126*(5), 2390–2412. https://doi.org/10.1121/1.3238250

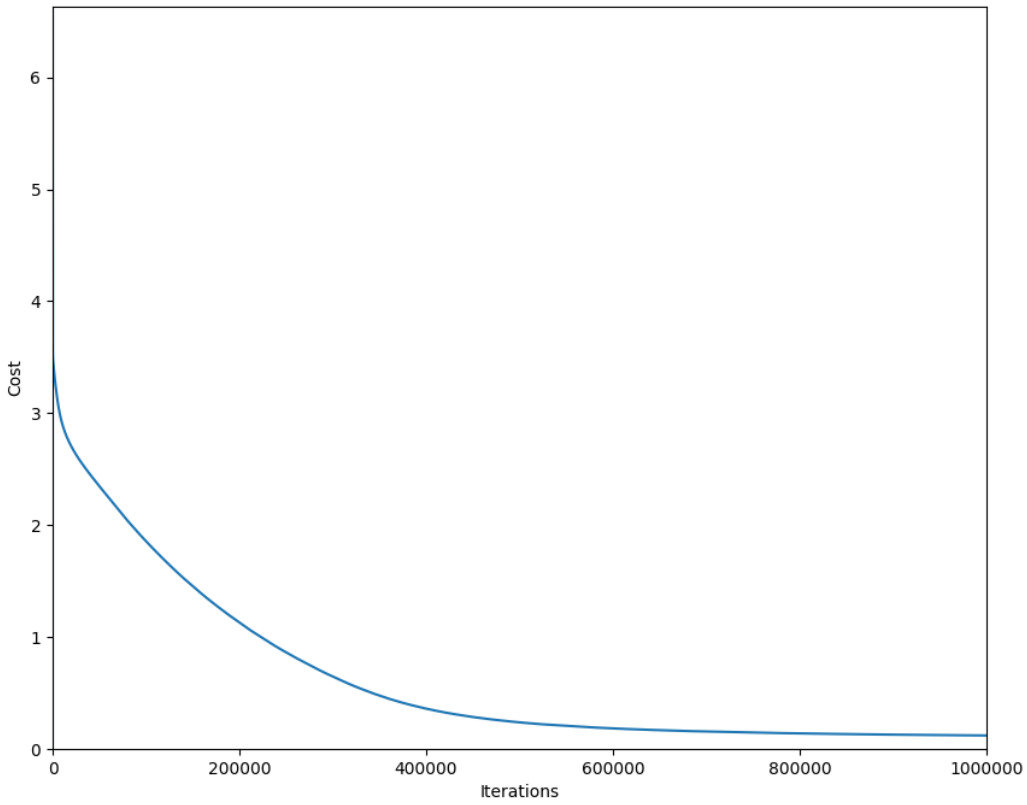
# 8 Appendix A: TCD-VoIP Neural Network Training Graphs




*TCD-VoIP Neural Network Training with 5k epochs*

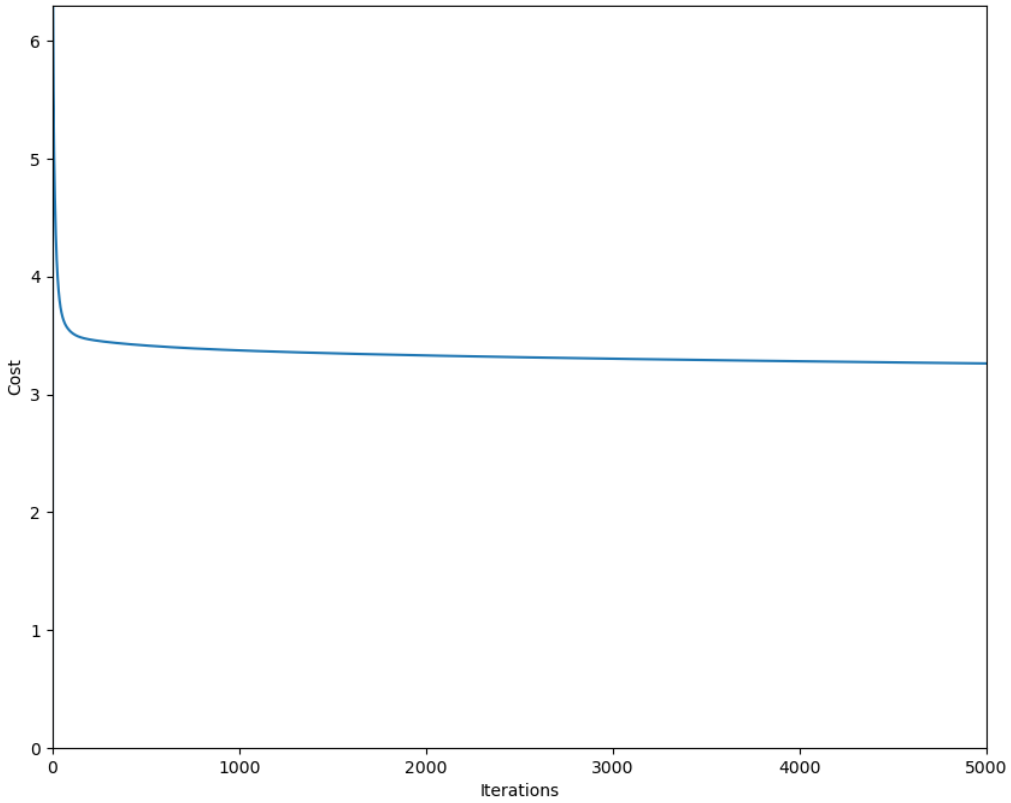



*TCD-VoIP Neural Network Training with 50k epochs*

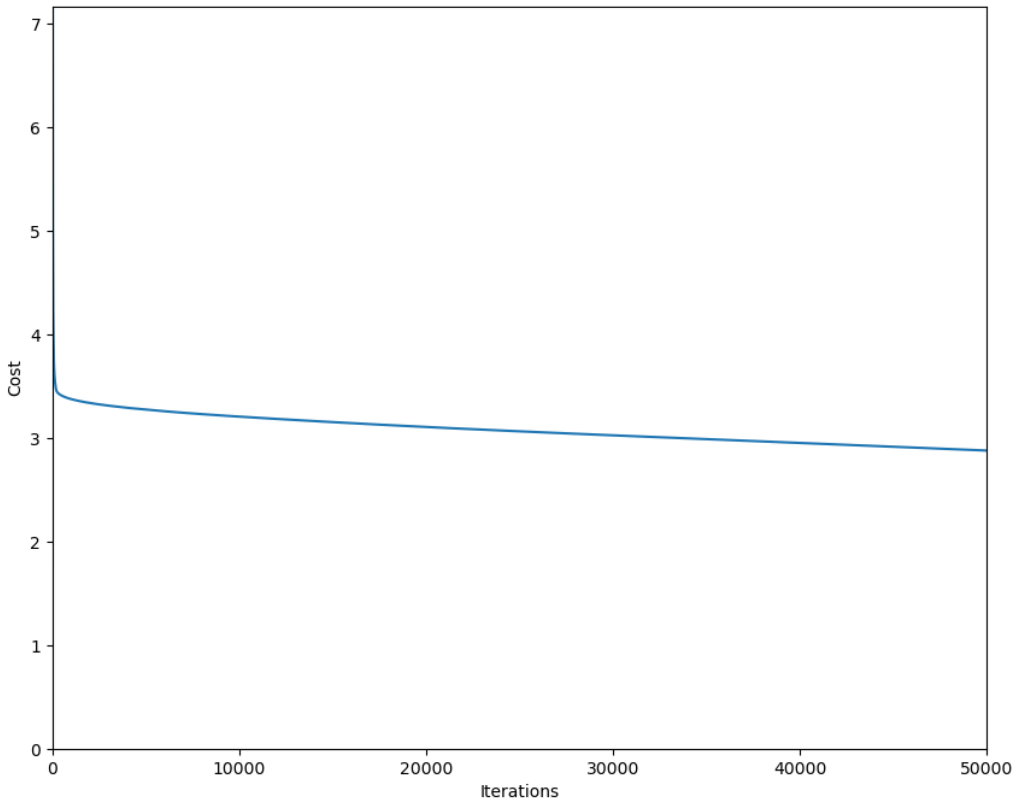



*TCD-VoIP Neural Network Training with 1M epochs*

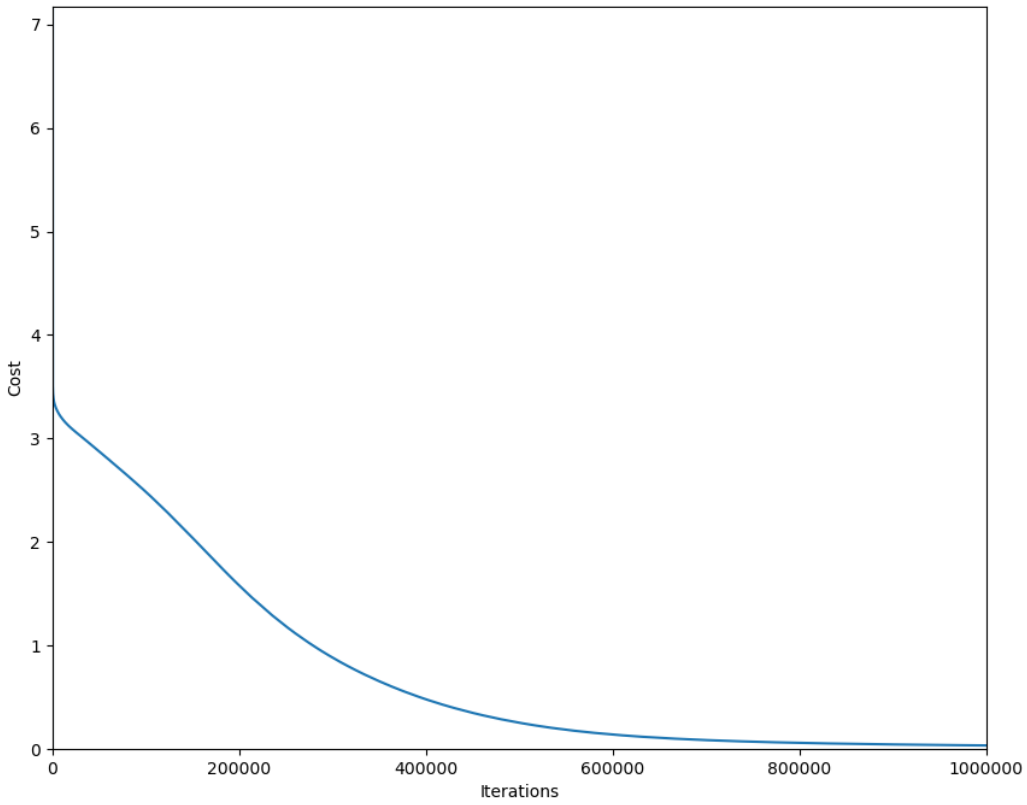
# 9 Appendix B: ITU-T P.Supp 23 Neural Network Training Graphs




*ITU-T P.Supp 23 Neural Network Training with 5k epochs*




*ITU-T P.Supp 23 Neural Network Training with 50k epochs*




*ITU-T P.Supp 23 Neural Network Training with 1M epochs*