Master of Science Thesis in Electrical Engineering
Department of Electrical Engineering, Linköping University, 2022

Evaluation of the Perceived Speech Quality for G729D and Opus

With Different Network Scenarios and an Implemented VoIP Application

Louise Almér

LINKÖPING UNIVERSITY

Master of Science Thesis in Electrical Engineering

Evaluation of the Perceived Speech Quality for G729D and Opus: With Different Network Scenarios and an Implemented VoIP Application

Louise Almér

LiTH-ISY-EX--22/5475--SE

Supervisor: **Harald Nautsch**
ISY, Linköpings Universitet
Carl Folkesson
Sectra Communication AB
Alexander Skoglund
Sectra Communication AB

Examiner: **Ingemar Ragnemalm**
ISY, Linköpings Universitet

Division of Information Coding
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping, Sweden

Abstract

Communication has always been a vital part of our society, and day-to-day communication is increasingly becoming more digital. VoIP (voice over IP) is used for real-time communication, and to be able to send the information over the internet must the speech be compressed to lower the number of bits needed for transmission. Codecs are used to compress the speech, or any other type of data transmitting over a network, which can introduce some noise if lossy compression is used. Depending on the bandwidth, bit rate, and codec used can distortion be minimized which would result in higher perceived speech quality.

In the thesis, two codecs, G729D and Opus, were tested and evaluated with two different objective perceive speech quality metrics, POLQA and PESQ. The codecs were also tested with different emulated network scenarios, 2G, 3G, 4G, satellite two-hop, and LAN. Furthermore, Opus was tested with and without VAD (voice activity detection) to see how VAD could affect the perceived speech quality. The different network scenarios did not impact the results of the evaluation, since the main difference between the network scenarios was latency, which POLQA and PESQ do not consider in the evaluation. Opus achieved a higher MOS-LQO (mean opinion score listening quality objective) than G729D. However, when VAD was enabled with Opus for a low bit rate, 8 kbit/s, the MOS-LQO was lower than without VAD.

Acknowledgments

I would like to thank my supervisors at Sectra Communication, Carl Folkesson, and Alexander Skoglund, for all the help and support during my thesis. Furthermore, I would like to thank the other employees at Sectra who took an interest in my thesis with great discussions which helped me gain new perspectives on my thesis. For the guidance and feedback on my thesis, I would like to extend my gratitude to my examiner, Ingemar Ragnemalm, and my supervisor, Harald Nautsch, at Linköping University. I would also like to thank my family and friends for the continuous support and the joy you give me. Furthermore, thanks to the thesis workers at Sectra Communication for the good times and interesting discussions. Finally, I want to thank Nina Argillander. I will always be grateful for all the support and laughter. You are missed.

Linköping, June 2022
Louise Almér

Contents

List of Figures

List of Tables

Notation

Abbreviation

Abbreviation	Description
ACR	Absolute Category Rating
ADC	Analog Digital Converter
HFE	High-Frequency Energy
ITU	International Telecommunication Union
MOS	Mean opinion Score
MOS-LQO	Mean Opinion Score Listening Quality Objective
MOS-LQS	Mean Opinion Score Listening Quality Subjective
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Assessment
PSTN	Public Switching Telephone Network
VAD	Voice activity detection
VOIP	Voice over IP
VQT	Voice Quality Test
QOE	Quality of Experience
QOS	Quality of Service

1 Introduction

In the following chapter, a brief introduction to the subject of the thesis is given. The research questions are presented, as well as the delimitation made to narrow the scope of the thesis.

1.1 Motivation

Communications have always been a vital part of our society, and day-to-day communication is increasingly becoming more digital. Furthermore, during the past couple of years, more people are working from home which has increased the demand for fast and reliable internet connection. The demand increases further when real-time communication, like telephone and/or video calls, transmits over the internet. The users expect the speech quality to be close to reality. However, distortion can occur which lower the perceived speech quality.

When several users transmit data at the same time, the channel capacity can be saturated. A solution is to use a codec (code-decode), which compresses the data to lower the number of bits needed to transmit the data with minimal quality deterioration [1]. With a lowered bit rate, the total occupied spaced on the channel for the data transmission is smaller than the original data size, and therefore decreases the chance of reaching the channel's maximum capacity. Since a codec compresses the data during transmission and thereafter decompress it at the receiver, some information can be lost, and some distortion might arise. Consequently, the choice of codec is important since different codecs will cause different amounts of distortion.

When applications such as WhatsApp, Skype, or Zoom are used for telephone calls, a technology called VoIP (voice-over-IP) [2] is used. VoIP, unlike the pub-

lic switching telephone network (PSTN), converts the analog signal to a digital signal, compresses it, and split it into small packets before transmitting the data over the internet. On the receiver side, the digital signal is decompressed, and the analog signal is estimated from the digital signal. To evaluate how a user perceives the quality of the estimated analog signal quality of experience (QoE) metrics are used. Usually, the signal is ranked with the mean opinion score (MOS) [3] which is between one (bad) to five (excellent).

The QoE is closely linked to the quality of service (QoS) [4]. QoS refers to the state of the network and different QoS metrics can be used to evaluate the transmission quality in terms of e.g., end-to-end packet delay, packet loss, and jitter. To produce accurate results for the QoE metrics, the QoS parameters should be fixed such that the audio codecs are evaluated with the same network conditions. If a real network would be used, the state of the network would likely change between the different tests since the traffic pattern can change during testing e.g., increased congestion. Reliable results can be achieved by using a network emulator that produces the same network scenario for the audio codec testing. Furthermore, multiple network scenarios, which are common for VoIP communication, should be used for the evaluation to achieve a reliable evaluation of the audio codecs.

1.2 Aim

The thesis aims to compare the perceived speech quality of two audio codecs, Opus and G729D. The codecs' performance will be evaluated with different network scenarios which are 2G, 3G, 4G, satellite, and satellite two-hop. Opus will be further evaluated with different bit rates, and with voice activity detection (VAD) enabled and disabled. A simplified VoIP application will be developed in C and used to stream speech files between two endpoints. A network emulator will be used to achieve reliable network conditions for the evaluation. Lastly, the perceived quality will be measured and evaluated with relevant QoE metrics.

1.3 Research Questions

The research questions that will be answered in the thesis are:

- What QoE metrics should be used to evaluate the perceived speech quality of G729D and Opus?
- How does Opus perform in comparison to G729D in different network scenarios regarding the perceived speech quality with narrowband bandwidth?
- How is the perceived speech quality affected when an 8 kbit/s and a 12 kbit/s bit rate on narrowband is used with Opus for different network scenarios?

- How is the perceived speech quality affected when a 16 kbit/s wideband is used for Opus for different network scenarios?
- Does voice activity detection (VAD) have any impact on the perceived speech quality for Opus with different bit rates?

1.4 Delimitations

For the thesis, two audio codecs will be compared and evaluated. G729D is evaluated since it is a well-established audio codec that is still used today for VoIP communication for low bit rates[1]. Opus was chosen to be compared with G729D, since it is a newer audio codec that has gained popularity since its release in 2012[2].

Opus can handle both mono and stereo audio, but since G729D only is adapted for mono audio, mono coupling will be used for Opus. Furthermore, Opus can operate on a wide range of bit rates and since the focus of the thesis is perceived speech quality will only three different bit rates for Opus be evaluated, 8 kbit/s and 12 kbit/s for narrowband, and 16 kbit/s for wideband. Narrowband and wideband are only evaluated with Opus, since these bandwidths ranges are most common to use for VoIP communication. Furthermore, Opus will be evaluated with and without voice activity detection and packet loss concealment, PLC, will be enabled for all evaluations for Opus.

The audio used for the experiment will consist of .pcm files instead of real-time audio. Therefore, the codecs will be tested and evaluated with the same audio files, which assures a fair result. The .pcm files contain speech from multiple male and female speakers. The network scenarios to evaluate the performance of the codec, 2G, 3G, 4G, satellite two-hop, and LAN, were chosen since these are the types of network scenarios to expect with VoIP communication.

1.5 Thesis Outline

In Chapter 2 relevant background information is presented to give a basic understanding of digital audio and transmission of audio. Also, details about audio compression, audio codecs and quality of experience metrics used for speech evaluation are described in the second chapter. Related work, regarding e.g. codecs and quality of experience metrics, is presented in Chapter 3. The implementation of the VoIP application, the audio datasets used and the test and evaluation setup are explained in Chapter 4. Chapter 5 consist of the results of the evaluation of the two codecs, the quality of experience metrics and datasets. The results are discussed in Chapter 6, with regard to the research questions and background.

[1] https://medium.com/@deborahlee_53049/the-advantages-of-using-g-729-codec-in-voip- accessed: 2022-05-03

[2] https://www.onsip.com/voip-resources/voip-fundamentals/opus-one-codec-to-rule-them-all, accessed: 2022-05-05

Some future work that can be implemented to continue the work is presented in Chapter 7 as well as some final conclusions.

2
Background

Background information is presented in the following chapter. First, some information about hearing, speech, and digital conversion, is shortly described. Second, an introduction to audio codecs with examples of different types of codecs are explained. Lastly, different types of quality of experience (QoE) metrics are described, with some examples of metrics used for perceived speech quality evaluation.

2.1 Sound Perception and Psychoacoustic Model

Sound traverse through the air, or any other medium, as a longitude wave. The wave is essentially a pressure variation that changes depending on how the air molecules move [5, Chapter 1]. When the sound wave reaches the ear, the sound pressure is converted to mechanical energy, which is converted to vertical waves and lastly as electrical impulses that are transmitted to the brain [5, Chapter 4].

A human's hearing change over time, but the average human can hear in the frequency range of 20 Hz to 20,000 Hz. However, the hearing range is also dependent on the loudness of the signal, which is measured in dB. Temporal and frequency masking are two types of sound masking that the human hearing is sensitive to. If a weaker sound follows a stronger sound, where both sounds have frequencies near the same range, the weaker signal is masked out and not audible due to temporal masking. Frequency masking occurs when two sounds have a similar frequency, such that one of the sounds is not audible when played at the same time [6, Chapter 10].

For modeling humans' perception of sound, psychoacoustic models are used. The models are constructed with information about the human hearing system and

the limitations of the human hearing. Psychoacoustic models are important to use for codec algorithms to achieve good perceived quality on the output signal [5, Chapter 4].

2.2 The Frequencies of the Human Speech

The frequency range of the human speech has been greatly researched. Mostly, the lowest number of frequencies needed for speech intelligibility have been studied, since most research is focused on bandwidth reduction for telephony. What these research have concluded is that a cut-off frequency at 4 kHz can be used and still produce decent speech [7]. The human voice is made of the fundamental frequency, F0, and the formants, F1, F2, etc., which are produced below 4 kHz. The fundamental frequency differs between female and male speech, which is caused by the anatomical and physiological differences between the genders, for instance the vocal tract length. The average fundamental frequency for males is around 120 Hz and for females is the fundamental frequency around 200 Hz [8].

In human speech, higher frequencies exist, usually between 5 and 10 kHz, which are known as High-Frequency Energy (HFE). HFE is mainly caused by the fricatives, a type of consonant. When developing the audio bandwidth for telephony, it was decided that HFE was not important for speech intelligibility and therefore was the frequency range limited to 3.4 kHz [9]. The fricatives' frequencies are different between the genders. HFE for female speech tends to be higher and contain more frequencies than for male speech [10]. It has been concluded, in more recent years, that the HFE does have an impact on speech intelligibility. Furthermore, speech with the HFE intact had better perceived speech quality, and had a positive impact on voice localization and speaker recognition [9].

2.3 Digital Representation

To represent analog audio digitally, an analog-to-digital converter (ADC) is used. The ADC converts the analog sound into voltage, which is thereafter measured at different sample points. Each of the measurements is mapped to a number, which is the digital representation of the analog signal. In Figure 2.1, an example of a digital representation of an analog signal can be seen.

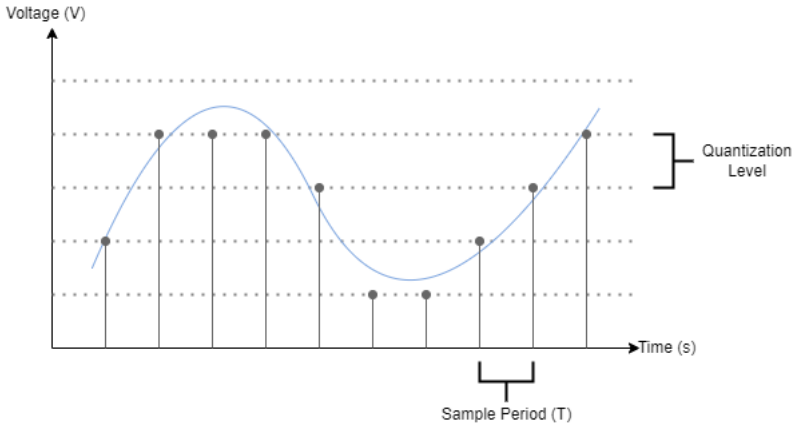

Figure 2.1: *Example of a digital signal (the gray dots) created from an analog signal (the blue line) by sampling and quantization*

The measurement mapping is called quantization and depending on how many quantization levels are used, the digital signal can be similar to the analog signal or very different. The number of quantization levels depends on how many bits are used to represent the signal. A typical .pcm file is sampled with 16 bits, which is equal to $2^{16} = 65536$ quantization levels. Besides the quantization, the sample rate is an important factor to create an accurate representation of the audio signal, where the sampling frequency $f_s = 1/$ sample period (T). According to the Nyquist theorem, the sample rate should be double the size of the maximum frequency in the audio signal. Aliasing may occur if the Nyquist theorem is not fulfilled, which will distort the digital signal [6, Chapter 10].

2.4 Audio Codecs

After the audio has been converted to a digital signal, codecs are used to further compress the signal to achieve a lower bit rate for the data transmission and to decompress the signal at the receiver. An audio codec contains a coder, which compresses the audio at the sender, and a decoder, that decompress the audio at the receiver. Depending on usage e.g., speech or music, different amount of data has to be compressed by the codec. The more frequencies in the signal, the more bits are needed to be able to reproduce it. Audio codecs are usually developed to transmit one (or few) types of audio signal that requires a certain amount of bandwidth. There are four defined audio bandwidth ranges: narrowband, wideband, super-wideband and fullband [11], which have been standardized by the International Telecommunication Union (ITU). The sample rates and bandwidth for each range can be seen in Table 2.1. Most common for telecommunication is narrowband or wideband bandwidth.

Table 2.1: *Bandwidth types with corresponding bandwidth range and sample rate*

Name	Audio Bandwidth	Sample Rate
Narrowband	300 - 3,400 Hz	8 kHz
Wideband	100 - 7,000 Hz	16 kHz
Super-wideband	50 - 14,000 Hz	32 kHz
Fullband	20 - 20,000 Hz	48 kHz

2.4.1 Speech Compression

There are two main categories of compression that exists, lossless and lossy. Lossless refers to compression where the original input is identical to the compressed and decompressed output signal. Since the output should be identical to the input, the compression cannot achieve as low bit rate as lossy compression. Therefore, lossy compression is usually used for audio compression. With lossy compression, more information is removed [1, Chapter 1]. For instance, frequencies humans cannot hear, or frequencies masked by temporal and frequency masking can be removed without any change of perceived speech quality.

Below are the three most common types of lossy speech codec techniques defined. A few codecs are presented which are of interest for the thesis, no specific waveform modeling codec is presented since that type was not used in the thesis.

Waveform Modeling Codec

A waveform modeling codec mimics the soundwave of the input signal and reconstruct the signal at the decoder. A benefit of waveform modeling codecs is that it is source independent, thus all input signals are coded equally. Therefore, waveform modeling codecs can handle input signals of varying noise level and speech characteristics [12, Chapter 7].

Vocoder

A vocoder (voice coder) is a type of source codec used as a speech codec, and consists of an analysis and synthesis method. The vocoder analyses the input signal and creates a model of the signal. The parameters of the model are thereafter transmitted instead of the actual signal. At the receiver, the parameters are used to synthetically create the input signal [13]. Since the parameters of the modelled signal is transmitted instead of an actual signal, a vocoder can achieve lower bit rates than a waveform modeling codec. However, the quality of vocoders are usually lower than waveform modeling codecs since the signal is estimated based on parameters instead of the soundwave [12, Chapter 7].

An example of a vocoder is G729D [14]. In 1996 G729 was first released as a recommendation from the International Telecommunication Union (ITU). G729 uses conjugate-structure algebraic-code-excited linear prediction (CS-ACELP) for

speech coding at 8 kbit/s with a speech frame of 10 ms. Two years later, in 1998 annex D was released, which performs coding at 6.4 kbit/s and was developed as a more flexible extension of G729. Since G729D has adjusted its parameters to lower the bit amount, the quality of the output signal for G729D is lower when compared with G729.

Another example of a vocoder codec is Enhanced Mixed Excitation Linear Predictive (MELPe) [15]. The codec was developed by the company Compandent between 1998 and 2001. MELPe is an enhanced version of MELP and can code speech at three different bit rates, 2400/1200/600 bit/s, and contains a Noise Pre-Processor (NPP) to reduce background noise. MELPe is used as a standard for the US Military (MIL-STD-3005) and NATO (STANAG 4591).

Hybrid Codec

A hybrid codec is the combination of waveform modeling codecs and vocoders. A synthetic model for human speech is used as well as data from the soundwave when compressing the input signal. The speech quality and bit rate of hybrid codecs are between waveform modeling codecs and vocoders [12, Chapter 7].

A hybrid codec that is widely used is Opus [16]. Opus audio codec was standardized as RFC 6716 by the Internet Engineering Task Force in 2012. Unlike the two previously presented audio codecs, MELPe and G729D, Opus was developed to be able to adapt such that the codec can be used for narrowband mono speech to fullband stereo music. Thus, Opus has a bit rate range from 6 kbit/s to 510 kbit/s. Opus is based on the SILK vocoder and the CELT waveform modeling codec. Opus has three different modes, SILK only, CELT only, and hybrid. The hybrid mode can only be used for bandwidths higher or equal to super-wideband. For low bandwidths, narrowband and wideband, SILK is used. The CELT only mode can be used for all bandwidths, narrowband to fullband, and is mostly used for music streaming.

2.4.2 Compression Enhancements

Codecs can have more functionalities than solely audio compression. Opus has a few different types of settings to enhance the compression [16], where two, packet loss concealment (PLC) and voice activity detection (VAD), are of interest for the thesis. These features are not an option for G729D. However, voice activity detection is available with G729D through annex F (G729F) [14]. In the following subsection is some further description given of PLC and VAD.

Packet Loss Concealment

PLC can be used to mask the effects of a lost packet. For VoIP, the latency of resending a packet will be noticeable, and therefore the better approach is to conceal the lost packet. When PLC enabled for the decoder, the signal can either be reconstructed from the previous received packet or interpolated between two speech frames, depending on the codec used. For Opus, both types of PLC can

be used, which depends on if SILK's or CELT's PLC is enabled. For speech transmitted on narrowband or wideband the SILK PLC is used, therefore the signal is reconstructed by the previous packet [16].

Voice Activity Detection

VAD can be used to lower the total number of bits needed for the data. At the sender, VAD detects whether the audio frame contain speech or non-speech. If speech is detected, the coder compresses the audio frame and transmit it. If the VAD does not detect any speech, the encoder will not be used, and a low bit rate comfort noise is usually transmitted instead. The reason for transmitting comfort noise instead of silence is to ensure the users that the VoIP call is still active. VAD, and the comfort noise generation, can affect the perceived speech quality of the signal if the signal-to-noise ratio is low. Some information can therefore be removed from the audio signal due to clipping when VAD is activated [17].

2.5 VoIP Communication, Networking and Network Emulation

Before VoIP, all telephone communication transmitted over the public switched telephone network (PSTN) with circuit switching. When circuit switching is used, each communication tunnel between two users reserves a space on the network until the communication is terminated. This means that a part of the channel is always occupied, even if nothing is sent between the users.

With VoIP, the communication is transmitted over the internet and therefore packet switching is used instead of circuit switching. Packet switching does not reserve any space on the channel for communication between users. The transmitted message is divided into small packets and sent over the network. If nothing is sent, the space on the channel can be used by other users on the network. The main benefit of packet switching is that more information can be sent over the network since no bandwidth is reserved for a specific communication tunnel. The drawback of packet switching is that collisions can occur between packets, and packets can get lost or delayed, which affects the quality of service (QoS) of the network [18, Chapter 4].

Different types of networks are prone to achieve higher or lower QoS e.g., a large public network with low bandwidth and a lot of traffic would most likely generate a low QoS value. The number of users affects the QoS, since more active users at the same time lead to more packets in transmission which increase the risk of delay, collision, etc. Since these factors cannot be controlled, network emulators can be used to mimic certain types of network scenarios during testing. With a network emulator, the QoS parameters are fixed for each emulated network such that each codec have the same precondition, and the evaluation would be solely on the QoE.

2.6 Quality of Experience

In the following section, some of the most common types of quality of experience (QoE) metrics used to evaluate the perceived sound quality are presented. Some examples of different metrics of each type are introduced and explained in more detail.

When measuring the quality of audio, the most common evaluation scale to use is the MOS scale [19], which is recommended by the ITU and can be seen in Table 2.2. The MOS scale is used for different types of evaluations e.g., subjective or objective. Therefore, different abbreviations are used, depending on the evaluation situation, which is explained further in ITU's recommendation *"Mean opinion score (mos) terminology"* [3].

Table 2.2: *The MOS scale and the definition for each number*

Score	Definition
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

2.6.1 Subjective and Objective Quality Assessment

Subjective listening tests have been well established to measure the perceived speech quality and were once the only evaluation method used to evaluate the perceived speech quality. When the perceived speech quality is evaluated subjectively, an audio sample is usually rated according to the ACR (absolute category rating) which is an integer scale from one (bad) to five (excellent). The MOS-LQS (mean opinion score listening quality subjective) can thereafter be calculated from multiple ACR ratings to estimate the overall perceived audio quality. The MOS-LQS scale is not restricted to integers, and the scale is between one (bad) and five (excellent) [19], as mentioned in the previous section.

However, since subjective assessment requires at least 20 subjects for the evaluation and the subjects must fulfill certain criteria and training to take part in the assessment [20], the subjective listening tests can be expensive in resources. The result of a subjective listening test can also yield differences between different subject groups, depending on the subject's values and the design of the subjective test [21]. For these reasons, new types of perceived speech quality assessment models have been developed to mimic the results of subjective listening tests.

Objective quality of experience metrics can be used to predict the perceived quality without using subjects for listening. Therefore, individuals' values are not applied to the listening score and the objective quality assessment can be used as

an average score of a subjective test group [21], usually mapped to the MOS-LQO (mean opinion score listening quality objective) scale. To mimic humans' perception of sound, psychoacoustic models are commonly used for objective metrics.

2.6.2 Intrusive and Non-Intrusive Models

There are two types of objective models used when evaluating the perceived speech quality of an audio signal, intrusive and non-intrusive models. An intrusive model, also known as full-reference or double-ended, compares the undistorted input signal with the degraded output signal to estimate the perceived speech quality. The non-intrusive model, also known as single-ended, uses the network parameters and the degraded output signal to estimate the perceived speech quality. The non-intrusive models yield a lower accuracy than intrusive models, which have been proven in [22–24]. Below are some intrusive and non-intrusive models presented which are recommended by the ITU when estimating the perceived speech quality.

Perceptual Objective Listening Quality Analysis

POLQA (Perceptual Objective Listening Quality Analysis) is a type of intrusive model which is recommended by the ITU with the name P.863 [21]. The model is developed to be used for objective listening tests, and therefore the perceived speech quality for two-way communication cannot be evaluated with the model. The perceived speech quality can be measured over the four different bandwidth ranges, seen in Table 2.1, by using two different modes, fullband mode, and narrowband mode.

Fullband mode can evaluate speech for all bandwidths, narrowband to fullband, and the degraded signal is rated with MOS-LQOf, where "f" stands for the fullband listening quality scale. The degraded signal in fullband mode must be externally upsampled to 48 kHz to be able to compare it to the 48 kHz sampled reference signal. Narrowband mode is adapted to solely narrowband and is rated with MOS-LQOn, where "n" stands for the narrowband listening quality scale. Consequently, the degraded signal should be sampled, or resampled, to 8 kHz to be able to be compared to the 8 kHz sampled reference signal. The different modes also contain different optimizations for the perceptional model used in POLQA. Therefore, due to different MOS-LQO scales and optimizations for the two modes, they cannot be compared to each other.

Even though the MOS-LQO scale is between 1 and 5, POLQA saturates the MOS-LQO for the signal when the score is ~4.5 for narrowband mode and ~4.8 for fullband mode. The MOS-LQO never reaches 5.0 since it represents the fact that some users in a subjective listening test would never rate the perceived speech quality a 5.0, even if the degraded and reference signal is the same.

Perceptual Evaluation of Speech Quality

The predecessor of POLQA, PESQ (Perceptual Evaluation of Speech Quality), is also an intrusive model which is mainly used for narrowband speech codec evaluation. Even though PESQ is still used as a recommendation from ITU under the name P.862 [25], there are some essential differences between PESQ and POLQA. POLQA considers bandwidth limiting, acoustic reverberations, and gain (signal level) differences when calculating the MOS-LQO, whilst PESQ is not adapted for these effects. The evaluation from POLQA has higher accuracy than PESQ according to the ITU [26], which is mostly due to POLQA being more adapted to modern telephony and was developed with a larger dataset of subjective quality test scores. Furthermore, the evaluation scale for PESQ differs from POLQA since it is between -0.5 and 4.5. For this reason, the raw PESQ score must be mapped to a MOS-LQO score. The function maps the PESQ values from -0.5 – 4.5 to 1.02 – 4.56 [27]. Similar to POLQA, PESQ does not score higher than 4.56, since the MOS-LQS retrieved from the subjective listening test never reaches 5.0. The mapping function can be seen in Equation 2.1, where x is the raw PESQ value and y is the mapped PESQ value.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.494*x+4.6607}} \tag{2.1}$$

PESQ also has a wideband extension called P.862.2 [28] that can be used for evaluation of the perceived speech quality of wideband audio. Similar to narrowband PESQ, the evaluation score is mapped to MOS-LQO with Equation 2.2, where x is the raw PESQ score and y is the mapped PESQ. The wideband Equation maps the raw PESQ values from 1.32 to 4.81.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.3669*x+3.8224}} \tag{2.2}$$

However, as stated in ITU's recommendation *"Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs"* [28], the output values from the mapped wideband PESQ should not be compared to the raw PESQ values nor the mapped narrowband PESQ values. The reason for this restriction is caused by the differences in the mapping functions.

P.563

A non-intrusive model for evaluating the perceived speech quality, that is currently recommended by ITU, is P.563 [29]. P.563 was developed to evaluate the perceived speech quality on narrowband and can be used for analog and digital signals. The model predicts listening scores by mapping the objective measurements to subjective measurements, like the MOS scale.

2.6.3 The E-Model

Another type of QoE model that can be used for evaluating perceived speech quality is the E-model, which is recommended by the ITU as G.107 [30]. The model is mostly used for the planning phase of a network and the QoE is calculated based on the QoS parameters. The model can predict the performance of some codecs, G729A and G723.1, regarding packet loss.

2.6.4 Data-Driven Methods

In real-life situations, the chance of having access to a clean reference signal for the evaluation with intrusive models is slim. Another drawback is that it is hard to evaluate the quality of a signal without a reference signal, and most non-intrusive models are built on assumptions about the audio signal which may not always correlate well with reality. For these reasons, data-driven models have been developed in recent years to evaluate the quality of speech with the use of e.g., tree-based regression and convolutional networks.

An example of a data-driven model is presented in X. Dong and D. S. Williamson's paper [22]. They developed a pyramid bidirectional long short-term memory network to predict the perceived speech quality ratings from crowdsourced evaluation in their paper. Another model called NISQA, based on CNN (convolutional neural network) was presented in G. Mittag and S. Möllers's paper [23]. Both models performed equally or better when compared to other established QoE metrics such as P.563 and POLQA.

3
Related Work

Related work to the thesis is presented in the following chapter. The related work is divided into different sections regarding the subject of the studies e.g., codecs, subjective or objective evaluation metrics.

3.1 Codecs

J. Skoglund and J.-M. Valin compared the quality of Opus at a low bit rate with and without neural speech generative models in their paper [31]. Two different models were used to improve the quality, WaveNet, and LPCNet. The models are used to re-synthesis the speech produced by the decoded signal, which can improve speech quality. In the experiment, a bit rate of 6 kbit/s was used on wideband bandwidth. From the subjective listening test could it be concluded that without the neural speech generative models, Opus produced a low MOS-LQS. The reason for Opus's low perceived speech quality for low bit rates is caused by the codec uses waveform matching when producing the decoded signal. J. Skoglund and J.-M. Valin states that the quality reduces greatly when the bit rate is below 10 kbit/s but with neural synthesis, the quality can be improved with low bit rates. For the thesis, bit rates above and below 10 kbit/s will be used. Therefore, it will be investigated if similar results are achieved for the thesis.

The performance of different speech codecs used for VoIP communications over satellite transmission was evaluated in F. Zampognaro et al. paper [32]. The evaluation metric used was PESQ and the codecs used for the experiment were GSM, Opus, G729, iLBC, and G723.1. From the evaluation could F. Zampognaro et al. conclude that Opus was the most suitable codec to use with VAD enabled, at a 10 kbit/s bit rate and a packet size of 80 ms. G729 was closely followed and was

deemed to be almost as good as Opus for VoIP over satellite transmissions. When the codecs are evaluated for the thesis, both Opus and G729D should perform well for satellite transmission. VAD was enabled for the best performance in F. Zampognaro et al. paper and therefore the effect of VAD will be investigated further in the thesis.

MELPe with two different bit rates, 1.2 kbit/s, and 2.4 kbit/s, was evaluated with PESQ and POLQA for Chinese speech in P. Souček and J. Holub's paper [33]. The audio files had different levels of distortion, therefore MELPe's perceived speech quality at different noise levels could be evaluated. The evaluation results from PESQ and POLQA were compared to a MOS-LQS using Pearson's correlation coefficient. From the results, it could be seen that the MOS-LQS was higher than the MOS-LQO generated by PESQ and POLQA. The authors conclude that the difference in scoring is because the participants were not as demanding as the objective metrics were regarding the quality, which is common for low-bit rate codecs. The perceived speech quality of the MELPe with Chinese speech gave a rather poor result, which could be the result of the parameters used in the metrics are mostly tuned for European languages and not Chinese. Lastly, P. Souček and J. Holub stated that PESQ was the most suitable QoE metric to use for these test conditions since the metric achieved a higher correlation to the MOS-LQS than POLQA for the Chinese language. The codecs that are evaluated in the thesis will not have as low bit rate as MELPe, therefore both POLQA and PESQ should be reasonable to use as QoE metrics.

3.2 Speech in Low Bit Rates

In J. Kaiser and T. Boril's paper [34] the vowel formant of the GSM ARM codec was evaluated using two different automatic formant extraction tools, Praat and VoiceSauce. In their experiment, five male and five female speakers were used with a variety of pitch (low/high) between the two genders. The three lowest formants, F1-F3, were extracted for evaluation. From J. Kaiser and T. Boril's results, a shift of formant frequencies could be observed for the codec for F2 and more present for F3. The shifted formants were mostly present for the female speakers and the authors concluded that the difference between fundamental frequency, F0, could be the cause since females have a higher F0 than males. The ARM codec was evaluated for both narrowband and wideband. Formant patterns were more intact for wideband than narrowband could J. Kaiser and T. Boril conclude from their results. For the thesis, neither Opus nor G729D shift the formants of the speech, instead the higher frequencies are filtered out.With shift or removal of frequencies, the signal is altered, which results in differences between the degraded signal and the reference signal. Consequently, the MOS-LQO from PESQ and POLQA can be affected by the removal of frequencies.

3.3 Evaluation of Subjective QoE Models

Laboratory and crowdsourcing subjective listening experiments were compared in T. Volk et al. article [35] regarding QoE evaluation for binaural playback in a teleconferencing scenario. T. Volk et al. investigated the possibility of replacing laboratory-based experiments with crowdsourcing. Laboratory experiments are quite costly since it demands a lot of resources such as time and money, which might not always be available when conducting listening tests. Furthermore, laboratory experiments do not reflect the real experienced quality of the subjects since the circumstances are closely monitored to follow the test protocols. In a real situation, subjects will have different types of hardware, such as headphone and computer brands, which will cause a different experience between users. The experiment was conducted with two different groups of crowdsourcing on different demographics, Germany and Great Britain/USA, and one laboratory experiment group in Germany. From the experiment could it be concluded that crowdsourcing was easier to perform with a larger and more diverse subject group. However, the results were not as reliable as the laboratory experiment since the intention and motivation of the crowdsourcing subjects can in some cases produce unreliable results. T. Volk et al. concluded that crowdsourcing could not be used instead of standardized laboratory experiments in their experiment. However, is it important to evaluate the QoE in a real-life scenario, and therefore crowdsourcing can be used to produce a more accurate description of the users' perception.

3.4 Evaluation of Objective QoE Models

In Y. Hu and P. C. Loizou's paper [36], a variety of different objective QoE metrics were evaluated. The evaluation consisted of measuring the correlation, p, between the MOS-LQO to scores received by subjective listening tests, MOS-LQS. There were 32 subjects that each scored the audio signal based on three scales: MOS (mean opinion score), SIG (signal distortion), and BAK (background intrusiveness). Each of these scales is rated between 1 and 5, where 1 is the lowest and 5 is the highest. segSNR, WWS, PESQ, LLR, IS, CEP, and fwSNRseg were the seven objective quality metrics evaluated in the study. Out of the seven different objective metrics, PESQ followed by LLR and fwSNRseg had the highest correlation with the subjective metrics. Hu and Loizou concluded that PESQ had the highest correlation, $p = 0.89$, where the other two top metrics scored $p = 0.85$. Since PESQ got the highest correlation with the subjective score, it is the most relevant to use in the thesis out of the presented QoE metrics.

A non-intrusive perceived speech quality assessment was presented in N. Nessler et al. paper [37]. The model was implemented with a deep neural network to estimate the quality according to the MOS scale. The model's result was compared to the intrusive metrics, POLQA and PESQ, and the non-intrusive metrics, NISQA and WAWENet. To evaluate the generated objective MOS, a subjective evaluated MOS, MOS-LQS, was used as a benchmark. The proposed model performed well

and came closest to the MOS-LQS. It was noticed during the experiment that POLQA was not well adapted to reverberations, which decrease the MOS significantly for samples containing a reverberation time (T60) longer than 0.5 seconds. In their paper, N. Nessler et al. state that the reason POLQA and PESQ generated worse results for reverberations is caused by the fact that these metrics were not developed for distortions such as reverberations. Instead, POLQA and PESQ were developed to evaluate distortions caused during the transmission, for instance, compression loss and codecs. Since the main goal of the thesis is to evaluate codecs, POLQA and PESQ are suitable evaluation metrics to use.

Three objective quality of experience metrics were compared in A. Hines et al. paper [24]. Two of the metrics were the ITU recommended POLQA and P.563 and the third metric was ViSQOL, which is also an intrusive metric. Subjective quality evaluation scores were used as a quality reference for the speech files. Pearson correlation coefficient, Spearman rank-order coefficients, and root-mean-square error were calculated between the subjective test score and the objective score for each metric. The speech files were divided into different subgroups: chop, clip, competing speaker, echo and noise. The subgroups were evaluated separately, all groups except clip and all subgroups together. A. Hines et al. could conclude that POLQA had a good correlation with the subjective test score for all subgroups, whilst ViSQOL had a problem with choppy speech. The non-intrusive metric, P.563, had the lowest correlation with the subjective test score and the authors stated that it should not be used for these types of evaluations. Since the undistorted input signal can be used in the evaluation part of the thesis, POLQA can be used for evaluation that should correlate well with what the perceived sound quality would have been in a subjective test.

POLQA and its predecessor PESQ were compared in J. G. Beerends et al. paper [38]. To evaluate the performance of the metrics, the correlation between the MOS-LQO and a 95 percent confidence interval of the MOS-LQS was calculated with the root-mean-square error (RMSE). J. G. Beerends et al. could conclude that POLQA outperformed PESQ for both narrowband and wideband quality estimation when comparing the results with the subjective measurement. Furthermore, POLQA was developed to account for the distortions that may occur in VoIP communications like time-stretching/compression, and reverberations and some codec noise which PESQ does not consider in its evaluation. Since both PESQ and POLQA are still recommended by the ITU to be used for objective listening evaluation, both metrics can be used for the evaluation of the codecs in the thesis.

3.5 Conversational QoE Models

A model for evaluating the perceived speech quality for conversational circumstances was developed in F. Koster and S. Möller's paper [39]. The authors began by creating a new type of subjective conversational test where the subjects rated three different aspects of a conversation: listening, speaking, and interaction. The

reason for this type of test setup was to locate what parts of a conversation have the most impact on the overall quality of the conversation. From the test results, it was concluded that the interaction part of the conversation had the highest impact factor. Furthermore, the interaction depended on the listening and speaking parts of the conversation. Delay during the interaction had a large impact on the overall MOS. Another distortion that affected the overall quality was echo during the speaking phase. From the result of the test, a multiple linear regression model was created to estimate the overall perceived conversational quality. When evaluating the proposed model with the results from the subjective test, a correlation of $p = 94$ was achieved. F. Koster and S. Möller concluded that their model should be tested with different conditions and the weighing of each of the three phases, listening, speaking, and interaction, had to be identified. For VoIP, the perceived conversational quality is important to measure the QoE,since if the delay is too high a conversation cannot continue even if the perceived speech quality is excellent. The proposed model by the authors will not be used for the thesis, since it is not available, nor should conversational metrics be used for measuring the perceived speech quality for different codecs.

4
Method

In this chapter, the method of the thesis is presented. The choices of evaluation models are motivated by the information presented in the background chapter, Chapter 2, and the related work chapter, Chapter 3. Lastly, the test and evaluation setup are explained, and the audio datasets used for the evaluation are presented.

4.1 Choice of Evaluation Metrics

When it comes to the quality of experience tests, most papers prefer the use of laboratory subjective listening tests. The listening environment can be controlled, and the perceived speech quality can be measured directly from the users. However, these tests can be quite costly in resources. The use of crowdsourced listening tests is a cheaper alternative. Consequently, these tests have other drawbacks where the listening environment and equipment are not controllable, and the result is not as reliable as the laboratory testing, as mentioned in [35]. For the thesis, subject testing was not used since these resources were not available.

Data-driven models have become popular for perceived speech quality evaluation. These models are usually developed as non-intrusive models, which is a good option to use in real-time streaming scenarios or other scenarios where the reference signal is not accessible. The mentioned data-driven models, [22] and [23], are open-sourced and therefore accessible to use for the evaluation. However, the test environment was located on a restricted network, which means that there was no access to the internet. Therefore, no unauthorized programs or scripts can be added, so data-driven models were not used for evaluation. Another non-intrusive model that was not used for the evaluation is the E-model.

Since the emulator emulated the same network conditions for the different codecs, the E-model would evaluate the codecs similarly since the network parameters are equal. Furthermore, the E-model has not been tested with G729D or Opus, which means that it is uncertain how well the evaluation result from the E-model would correspond with the actual QoE for the codecs.

Since the input signal was available to use for the evaluation, intrusive metrics were used. The two main intrusive evaluation metrics that are used and recommended by the ITU for perceived speech quality evaluation on narrowband are POLQA and PESQ. As mentioned in [26], POLQA is more adapted to modern telephony and is assumed to achieve a more accurate result than PESQ. However, PESQ can be used and compared with POLQA to evaluate the performance for narrowband speech, but the accuracy of PESQ should be considered. Furthermore, PESQ and POLQA were developed for evaluating noise created from e.g. codecs, as mentioned in [37], which further increase their relevance as evaluation metrics. POLQA and PESQ were both available for the thesis and therefore were used for the perceived speech quality evaluation. For the evaluation of narrowband speech was the narrowband mode of POLQA used since the evaluation results can be compared to the evaluation results of PESQ [26]. For wideband speech with Opus, the fullband mode of POLQA [21] was used, as well as the wideband extension of PESQ [28].

4.2 Emulated Network Scenarios

The network scenarios used in the emulation were 2G, 3G, 4G, satellite two-hop, and LAN. The reason for using multiple different network scenarios for the evaluation was to evaluate the overall performance of the codecs. The emulator that was used for the thesis was iTrinegy[3]. The specifications of the emulated network scenarios are presented in Table 4.1. The most significant differences between the scenarios were the link speed and latency of the networks. The network emulator and network scenarios were provided by Sectra.

Table 4.1: *Specifications of the different network scenarios for the emulation*

Specific settings for emulated network scenarios			
Network Scenario	Link Speed (bps)	Min-Max Latency	Loss
2G	60,000	250 – 450 ms	2 %
3G	300,000	50 – 100 ms	1 %
4G	2,000,000	20 – 40 ms	1 %
Satellite two-hop	256,000	1,200 – 1,600 ms	1 %
LAN	100,000,000	5 – 9 ms	0.1 %

When the perceived speech quality for different codecs is evaluated, no conversational factors, such as latency, are measured. Therefore, one of the main differ-

[3]https://itrinegy.com/, accessed: 2022-03-23

ences between the network scenarios, latency, was not considered for the evaluation with POLQA and PESQ. This caused the results between the different network scenarios to be quite similar.

4.3 Codec Settings

The settings used for the codecs are presented below. Opus has more parameters that can be adjusted, therefore the main part of the section is the motivation of the chosen Opus parameters.

4.3.1 G729D

The settings used for G729D are fixed, and therefore they cannot, and should not, be changed. According to the documentation [14], the sample rate used for the codec should be 8 kHz with a 6.4 kbit/s bit rate, and the speech frame should be 10 ms.

4.3.2 Opus

Opus version 1.3.1[4] was used in the thesis. From the documentation [16], the recommended bit rates to use for speech compression are between 8-12 kbit/s for narrowband speech and 16-20 kbit/s interval for wideband speech. Therefore, Opus was evaluated with 8 kbit/s and 12 kbit/s bit rates for narrowband and a bit rate of 16 kbit/s for wideband. According to [31], the perceived speech quality with Opus degraded quickly below 10 kbit/s which made it interesting to evaluate the two chosen bit rates for narrowband. One bit rate was evaluated with wideband speech since the focus of the thesis was to evaluate the perceived speech quality between G729D and Opus on narrowband. Furthermore, Opus supports both mono and stereo coupling, but for the thesis was only mono coupling used since G729D operates with mono audio.

As mentioned in Section 2.4.1, Opus can operate with three different modes. Since the evaluation was for speech on low bandwidths, the SILK only mode was used in the codec. Opus has an internal sample rate for the compression and with the SILK only mode the internal sample rate is double the amount of the audio bandwidth. Consequently, since the speech files were externally sampled with 8 kHz for narrowband and 16 kHz for wideband, the audio bandwidths are 4 kHz respectively 8 kHz, due to the Nyquist theorem. Therefore, the internal sample rate is the same as the external sample rate, 8 kHz for narrowband speech and 16 kHz for wideband speech.

Lastly, Opus can operate with several speech frame lengths, from 10 ms to 60 ms with the SILK only mode. For the thesis, a 20 ms speech frame was used, since it is recommended as a "good choice" from the Opus documentation [16].

[4] https://opus-codec.org/release/stable/2019/04/12/libopus-1_3_1.html, accessed: 2022-04-26

Furthermore, PLC and VAD were used for the evaluation. PLC was enabled for all tests and the effect of VAD was investigated since it can impact the perceived speech quality, as mentioned in Section 2.4.2 and the paper by J. Skoglund and J.-M Valin [31].

4.4 Implementation

To test the perceived speech quality of the codecs, the speech signal must traverse through a network to emulate the VoIP communication and thereafter be compared with a reference signal. The library PJSIP was used in the thesis such that a call can be established between two endpoints. The library includes signal protocols, and therefore the library can be used to send and receive packets over the network. PJSIP is written in C and contains multiple APIs that can be used to manipulate the communication stream. Other open-sourced libraries exist that could have been used instead of PJSIP, e.g. Linphone. However, it was concluded that PJSIP had a more thorough documentation than Linphone and therefore was PJSIP chosen for the implementation.

The implementation of the application was divided into a few steps. First, the PJSIP library was used with an already implemented codec, G711, to setup a VoIP stream between two endpoints. Since the main focus of the thesis was not to build a VoIP application from scratch, one of PJSIP's example files, streamutil.c, was used as a foundation for the implementation. The file contained low-level APIs of PJSIP: PJMEDIA and PJLIB. Other, more high-level APIs within PJSIP were considered e.g., PJSUA, but low-level APIs were used to gain a better understanding of the PJSIP library.

The second step was to implement the third-party codecs. Opus was added first since it already had some implementation in PJSIP. For Opus, PJSIP had to be compiled with the added codec to register Opus to the codec manager such that it can be used during runtime. After Opus was implemented, the integration of G729D into PJSIP was the next step.

There are some guidelines[5] to follow when adding a new third-party codec in PJSIP. To register G729D to the codec manager a codec factory was created for G729D and the third-party codec was linked with the PJSIP library. The encoding and decoding functions were implemented similarly to the coder and decoder main files from the G729D library.

A third codec, MELPe, was planned to be integrated into PJSIP. However, due to time limitations, MELPe was not successfully implemented in the PJSIP library in time for the test and evaluation phase of the thesis. MELPe was meant to be implemented with an 8 kHz sample rate and a fixed bit rate of 2.4 kbit/s. Since the codec has a speech interval of 22.5 ms, two frames would have been sent per packet due to PJSIP's constraint of only using integer-based speech intervals,

[5] PJSIP codec framework `https://www.pjsip.org/pjmedia/docs/html/group__PJMEDIA__CODEC.htm`, accessed: 2022-03-14.

which would trick PJSIP into accepting MELPe as a codec with a speech interval of 45 ms.

4.4.1 Audio Datasets Used for the Evaluation

Since PESQ and POLQA were used for the evaluation of the codecs, some constraints exist for the input audio. According to the ITU application guide for POLQA [26], the reference signal should be at most 12 seconds long, with at least three seconds of speech and at least 500 ms of silence between active speech duration. At most, the total amount of active speech should be six seconds.

The audio files used for the emulation tests consisted of three different sets of audio files. The first set contained one female and one male speaker that spoke two short sentences in English, sampled at 8 kHz. The second set contained three female and three male speakers. The audio files were sampled at 8 kHz and the speakers did not speak any type of language but produced a sound that is well suited for testing the performance of the codecs with a variety of frequencies. The third and last dataset contained the same speakers as the first dataset, except that the audio was sampled at 16 kHz to evaluate Opus's perceived speech quality for wideband speech.

The datasets, with their corresponding language, number of speakers, and sample rates are seen in Table 4.2. All speech files were provided through GL communication's VQuad Probe[6], which was used for the evaluation and can evaluate the perceived speech quality with both POLQA and PESQ.

Table 4.2: *Details about each audio dataset used for the evaluation tests, in regard to number of speakers, language and sample rate*

Name	Speakers	Language	Sample Rate
Dataset 1	One female & one male	English	8 kHz
Dataset 2	Three females & three males	Gibberish	8 kHz
Dataset 3	One female & one male	English	16 kHz & 48 kHz

For each of the datasets, the male and female speakers were tested and evaluated separately to investigate if there were any differences in the perceived speech quality between the genders. As mentioned in the paper by J. Kaiser and T. Boril [34] and in Section 2.1, there is a difference in frequencies for female and male speech. Furthermore, for the POLQA evaluation for wideband speech must the reference signal be sampled at 48 kHz for the evaluation. Therefore, two sample rates are used for the third dataset, as mentioned in Section 2.6.2.

[6] `https://www.gl.com/vquad.html`, accessed: 2022-03-25

4.5 Emulation Setup

To test and evaluate the codecs, GL communications VQuad Probe was used with the Voice Quality Test (VQT), which was provided by Sectra. Two VQuad Probes were used, one for POLQA and one for PESQ evaluation, since the PESQ and POLQA licenses were on two different VQuad machines.

The VQuad contained the VQT program and was connected to a sound card. The sound card was further connected to the computer that ran the VoIP stream. The computer was also connected with a network cable to the network emulator such that the different network scenarios could be tested with the codecs. Therefore, the sound could traverse from the VQuad through the VoIP stream and network emulator, back to the VQuad, where the reference and the degraded signal could be compared with the VQT. The emulation setup can be seen in Figure 4.1. For each measurement in the experiment, the audio was streamed for ~15 minutes to confirm that the result of the evaluation would be trustworthy.

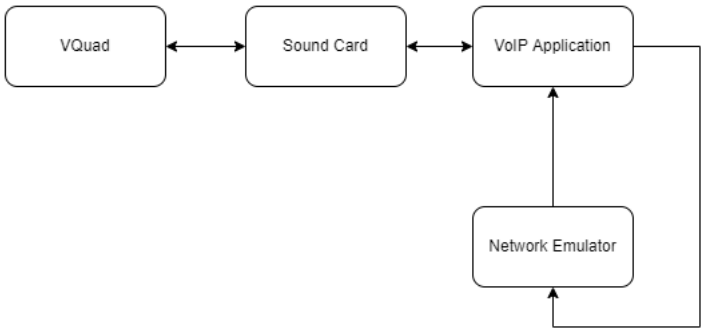

Figure 4.1: *Network emulation setup*

5
Results

In the following chapter, the results from the evaluation of the audio codecs are presented. The subchapters are divided based on the two codecs used, G729D and Opus. G729D was tested with the first two datasets since the codec is not adapted for wideband speech and therefore cannot be tested with the third dataset. Opus was evaluated with all three datasets since Opus can handle both narrowband and wideband bandwidth. As mentioned in Section 4.4.1, the male and female speakers have been evaluated separately. In the figures, the MOS-LQO for the female speakers is represented with blue dots and the MOS-LQO for the male speakers is represented with red dots.

5.1 Evaluation Results with G729D

In the following two figures, Figure 5.1 and Figure 5.2, the result of the POLQA evaluation with G729D can be seen. In Figure 5.1 the evaluation results for dataset 1 can be seen and the evaluation results from the dataset 2 can be seen in Figure 5.2. Thereafter, the next two figures, Figure 5.3 and Figure 5.4, contains the PESQ evaluation of dataset 1 and dataset 2 with G729D. The sample rate used for the audio signals were 8 kHz and the bit rate used was 6.4 kbit/s.

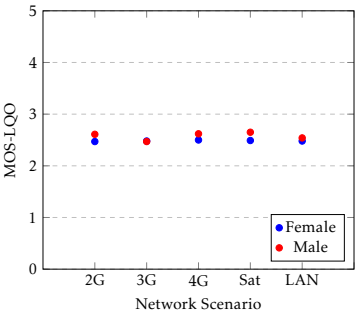

Figure 5.1: *POLQA evaluation of dataset 1 with G729D, 8 kHz sample rate and 6.4 kbit/s bit rate*

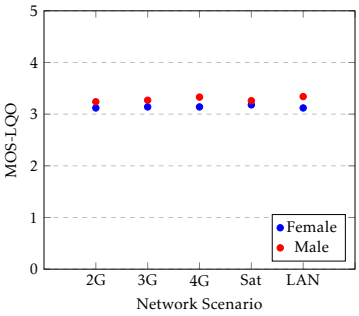

Figure 5.2: *POLQA evaluation of dataset 2 with G729D, 8 kHz sample rate and 6.4 kbit/s bit rate*

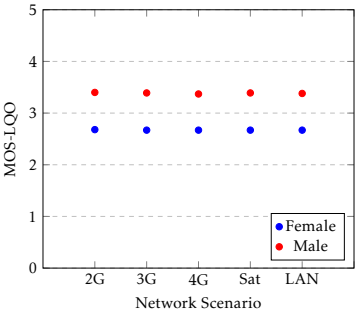

Figure 5.3: *PESQ evaluation of dataset 1 with G729D, 8 kHz sample rate and 6.4 kbit/s bit rate*

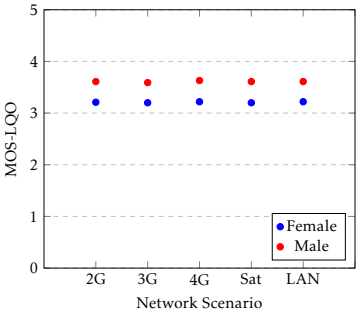

Figure 5.4: *PESQ evaluation of dataset 2 with G729D, 8 kHz sample rate and 6.4 kbit/s bit rate*

5.2 Opus

Opus was evaluated with VAD enabled and disabled. Furthermore, different bit rates were used for narrowband, 8 kbit/s and 12 kbit/s, to investigate the effect of different bit rates on the perceived speech quality. For wideband, one bit rate, 16 kbit/s, was used. The result of the evaluation is divided into three subsections, Section 5.2.1 to Section 5.2.3, based on the dataset used.

5.2.1 Evaluation Results with Dataset 1

The following four figures, Figure 5.5 – Figure 5.8, contains the results of the POLQA evaluation based on dataset 1. In Figure 5.5 Opus was evaluated with

voice activity detection (VAD) enabled and a bit rate of 8 kbit/s was used. The result of Opus with the same bit rate but without VAD enabled can be seen in Figure 5.6. The result of the evaluation of the higher bit rate, 12 kbit/s, for narrowband can be seen in Figures 5.7 and 5.8 with VAD enabled respectively disabled.

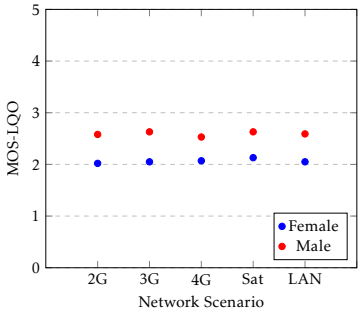

Figure 5.5: *POLQA evaluation of dataset 1 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD enabled*

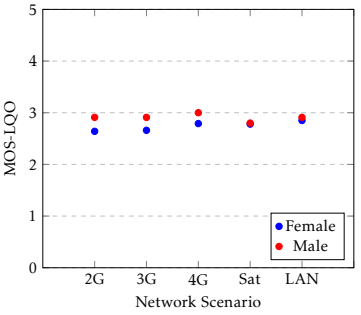

Figure 5.6: *POLQA evaluation of dataset 1 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD disabled*

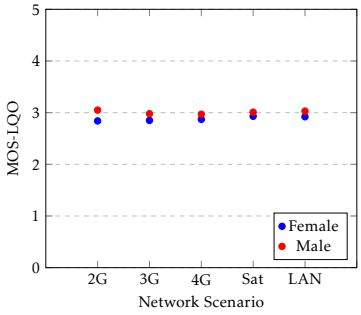

Figure 5.7: *POLQA evaluation of dataset 1 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD enabled*

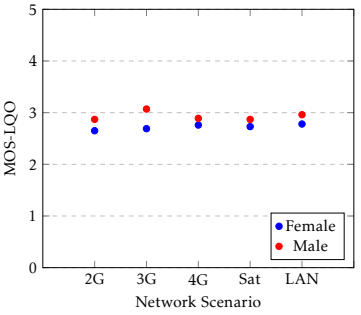

Figure 5.8: *POLQA evaluation of dataset 1 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD disabled*

In the next four figures, Figure 5.9 – Figure 5.12, the result of the PESQ evaluation with dataset 1 can be seen. In Figure 5.9 Opus was evaluated with VAD enabled and a bit rate of 8 kbit/s was used. The result with the same bit rate but without VAD enabled can be seen in Figure 5.10. The result of using 12 kbit/s bit rate can be seen in Figure 5.11 and Figure 5.12 with VAD enabled, respectively disabled.

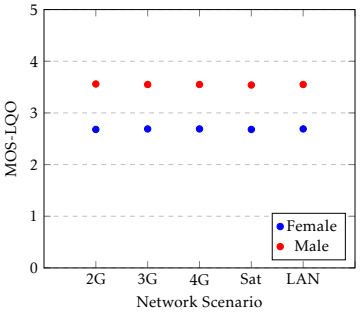

Figure 5.9: *PESQ evaluation of dataset 1 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD enabled*

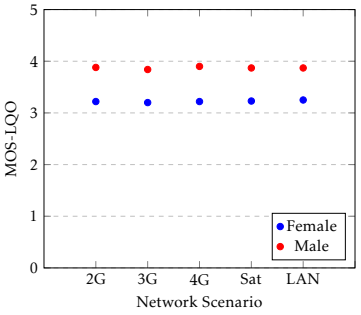

Figure 5.10: *PESQ evaluation of dataset 1 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD disabled*

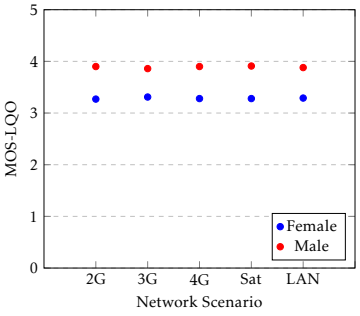

Figure 5.11: *PESQ evaluation of dataset 1 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD enabled*

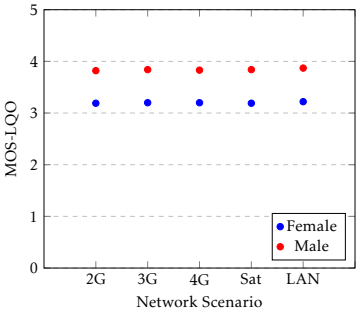

Figure 5.12: *PESQ evaluation of dataset 1 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD disabled*

5.2.2 Evaluation Results with Dataset 2

In the following four figures, Figure 5.13 – Figure 5.16, the result of the POLQA evaluation based on dataset 2 are seen. In Figure 5.13 Opus was evaluated with voice activity detection (VAD) enabled and a bit rate of 8 kbit/s was used. The result with the same bit rate but without VAD enabled can be seen in Figure 5.14. The result of the evaluation of the higher bit rate, 12 kbit/s, for narrowband is presented Figure 5.15 and Figure 5.16 with VAD enabled respectively disabled.

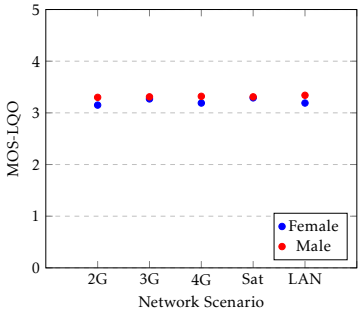

Figure 5.13: *POLQA evaluation of dataset 2 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD enabled*

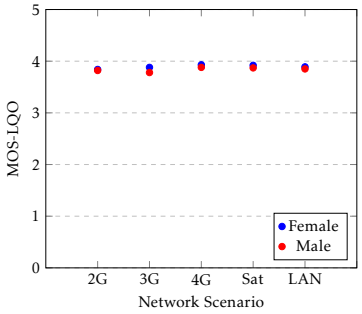

Figure 5.14: *POLQA evaluation of dataset 2 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD disabled*

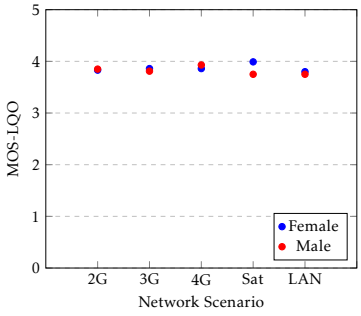

Figure 5.15: *POLQA evaluation of dataset 2 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD enabled*

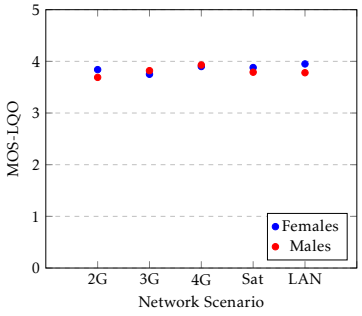

Figure 5.16: *POLQA evaluation of dataset 2 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD disabled*

The next four figures, Figure 5.17 – Figure 5.20, contains the result of the PESQ evaluation with dataset 2. In Figure 5.17 Opus was evaluated with VAD enabled and a bit rate of 8 kbit/s. The result with the same bit rate but without VAD enabled can be seen in Figure 5.18. The evaluation result of using a bit rate of 12 kbit/s can be seen in Figure 5.19 and Figure 5.20 with VAD enabled, respectively disabled.

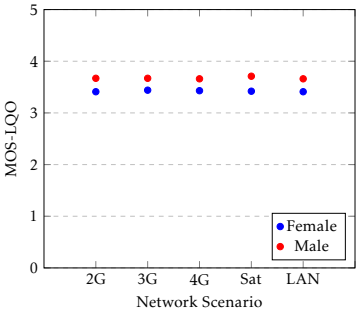

Figure 5.17: *PESQ evaluation of dataset 2 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD enabled*

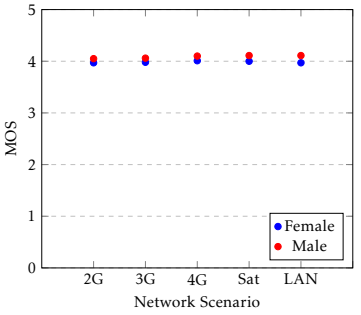

Figure 5.18: *PESQ evaluation of dataset 2 with Opus, 8 kHz sample rate, 8 kbit/s bit rate and VAD disabled*

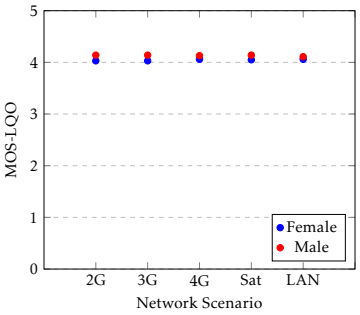

Figure 5.19: *PESQ evaluation of dataset 2 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD enabled*

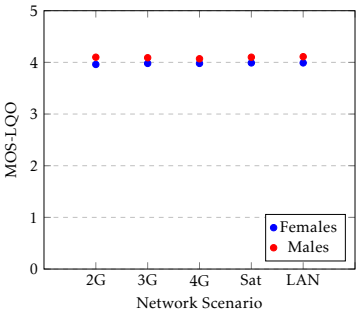

Figure 5.20: *PESQ evaluation of dataset 2 with Opus, 8 kHz sample rate, 12 kbit/s bit rate and VAD disabled*

5.2.3 Evaluation Results with Dataset 3

In the following four figures, Figure 5.21 – Figure 5.24, the evaluation based on dataset 3 with Opus can be seen. A constant bit rate of 16 kbit/s was used for the tests. The POLQA evaluation can be seen in Figure 5.21 with VAD enabled and in Figure 5.22 with VAD disabled. The PESQ evaluation can be seen in Figure 5.23 with VAD enabled and in Figure 5.24 with VAD disabled.

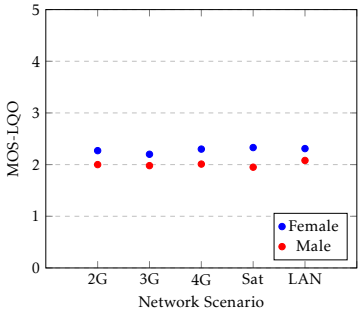

Figure 5.21: *POLQA evaluation of dataset 3 with Opus, 16 kHz sample rate, 16 kbit/s bit rate and VAD enabled*

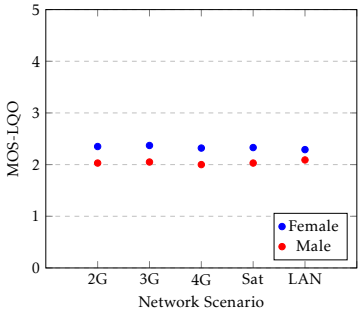

Figure 5.22: *POLQA evaluation of dataset 3 with Opus, 16 kHz sample rate, 16 kbit/s bit rate and VAD disabled*

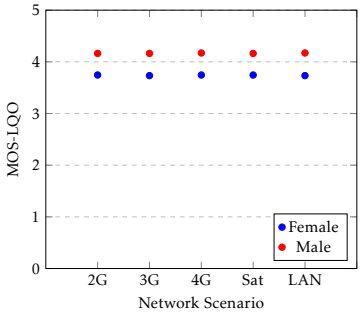

Figure 5.23: *PESQ evaluation of dataset 3 with Opus, 16 kHz sample rate, 16 kbit/s bit rate and VAD enabled*

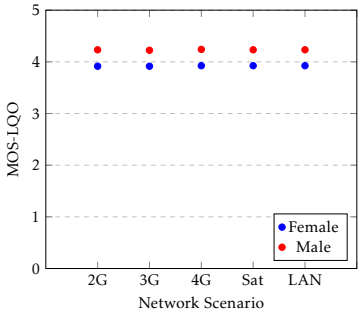

Figure 5.24: *PESQ evaluation of dataset 3 with Opus, 16 kHz sample rate, 16 kbit/s bit rate and VAD disabled*

6 Discussion

In this chapter, the result of the evaluation, seen in Chapter 5, is discussed based on the audio datasets, codec, and evaluation model used. The impact of the network scenarios and voice activity detection is also discussed. Thereafter, the method of the thesis is motivated and criticized, as well as the sources used. Lastly, a wider context for the thesis is presented.

6.1 Evaluation Results

The result presented in Chapter 5 is discussed in the following subsections. G729D was only tested for dataset 1 and dataset 2 since G729D does not support wideband speech.

6.1.1 Perceived Speech Quality with Dataset 1

Before the test and evaluation started with dataset 1, an observation was made. The sound level of the audio file containing the female voice was lower than the audio file with the male voice. To counter the gain difference, the sound level of the female voice was adjusted to match the sound level of the male voice. Without the sound level adjustment, a lower MOS-LQO was observed for the female speaker, which was caused by a low SNR (signal-to-noise ratio). With a low SNR, the noise makes it harder to distinguish the voice in the signal from the background noise.

From the result of the POLQA evaluation with Opus, seen in Figure 5.6 – Figure 5.8, a small MOS-LQO difference exist between the two speakers even after the gain adjustment, around 0.15, which indicates that POLQA was not completely unaffected by the sound level difference. Since the gain was increased for the

audio of the female voice, the noise in the audio file was also increased. The difference between the reference signal and the degraded signal might not be noted by humans. However, when analyzing the signals with PESQ and POLQA, small changes can cause a decreased MOS-LQO with e.g., higher noise levels. The most significant difference can be seen in Figure 5.5 which implies that voice activity detection, VAD, had an impact on the perceived speech quality for low bandwidths, which is further discussed in Section 6.1.7. The difference between the two speakers was more defined for the evaluation of the perceived speech quality with PESQ, as seen in Figure 5.9 – Figure 5.12. The cause for the large difference between the speakers, as mentioned in ITU's recommendation [26], could be that PESQ did not consider the changed sound level for the degraded signal for the female speech evaluation. Therefore, the female speech was scored with a lower MOS-LQO than the male speaker.

The perceived speech quality with Opus was not affected when a bit rate of 12 kbit/s was used instead of 8 kbit/s, with VAD disabled. Therefore, the increased bit rate does not correlate with the perceived speech quality. These observations can be seen for both POLQA and PESQ evaluations for dataset 1.

For G729D, the evaluation result with POLQA can be seen in Figure 5.1, and for PESQ is the evaluation result presented in Figure 5.3. Like Opus, a larger difference between the female and male speaker exists when evaluating the perceived audio with PESQ than with POLQA. The perceived speech quality with G729D had an overall lower MOS-LQO than Opus for dataset 1.

An interesting observation when comparing the figures is that it seems that PESQ overestimates the male speaker since the MOS-LQO of the female speaker with PESQ have similar values as the MOS-LQO for both the male and female speaker when evaluating with POLQA. These observations can be seen for both Opus and G729D. The difference between the female and male speech can be explained by that the perception model used in PESQ represents male speech better than female speech, which would explain why the female speech was evaluated with a lower MOS-LQO. The difference in MOS-LQO can also be explained by the gain adjustment made for the female speech.

6.1.2 Perceived Speech Quality with Dataset 2

The results of the evaluation for the second dataset, which contained the gibberish speakers, can be seen in Section 5.2.2 with Opus. In Figure 5.2 and Figure 5.4 the result of the evaluation can be seen when G729D was used as the codec.

From Section 5.2.2 it can be seen that both PESQ and POLQA rated the perceived speech quality quite similarly. As seen in Figure 5.14 – Figure 5.16, there was no defined difference between the male and female speakers with the POLQA evaluation, which would indicate that the files did not contain a noticeable gain difference. The average perceived speech quality of the POLQA evaluation for the tests was ~3.9. Seen in Figure 5.13, the MOS-LQO was lower, about ~3.1 which would indicate that VAD also interfered with the perceived speech quality

for dataset 2.

The PESQ evaluation of Opus had a slightly higher overall evaluation score than POLQA for the second dataset, ~4.1 on the MOS-LQO scale, which can be seen in Figure 5.18 – Figure 5.20. From these figures, a trend was observed that the perceived speech quality for the female speakers was slightly lower than for the male speakers. As mentioned in previous section, the difference between the female and male speech can be explained by the perception model used in PESQ. A similar result can be seen in Figure 5.13 and Figure 5.17, where the perceived speech quality was lower and a more defined difference exists between the female and male speakers when VAD was enabled. These results would further indicate that VAD had an impact on the perceived speech quality. Furthermore, when comparing the PESQ evaluation in Figure 5.18 with the POLQA evaluation with the same test parameters in Figure 5.14 it seems that PESQ overestimates the male speech since the female speech scores similar to the POLQA evaluation of both male and female speech. These results would further indicate that the perception model used in PESQ is more adapted for male speech.

Regarding the different bit rates used for the evaluation, a similar result to dataset 1 was seen for dataset 2. The MOS-LQO was not affected when a bit rate of 12 kbit/s was used instead of an 8 kbit/s bit rate, which can be seen for both the POLQA and PESQ evaluation. VAD did however affect the MOS-LQO and is further discussed in Section 6.1.7.

The POLQA evaluation gave a more cohesive evaluation between the female speakers and male speakers than the PESQ evaluation when G729D was used, as seen in Figure 5.2 and Figure 5.4. Like dataset 1, PESQ seems to overestimate the male speech, since the female speech evaluated with PESQ has similar MOS-LQO as the POLQA evaluation for both male and female speakers. When comparing the overall MOS-LQO between G729D and Opus for dataset 2, it can be concluded that Opus generates a higher MOS-LQO score than G729D, except when VAD was enabled with an 8 kbit/s bit rate.

6.1.3 Perceived Speech Quality with Dataset 3

As explained in ITU's recommendation [26], [27] and [28], the result obtained from the wideband evaluation cannot be compared with the narrowband evaluation. PESQ uses a different mapping function, see Equation 2.1 and Equation 2.2, and POLQA's two different modes, fullband mode and narrowband mode, uses different optimizations for the perception model, as explained in Section 2.6.2. Furthermore, the POLQA fullband mode and PESQ evaluation for wideband speech cannot be compared, since the MOS-LQO scales are different.

The perceived speech quality evaluation with POLQA can be seen in Figure 5.22 and Figure 5.21. For the POLQA fullband mode evaluation, the reference signal can contain frequencies up to 20 kHz and the degraded signal can only contain frequencies up to 8 kHz. Consequently, the MOS-LQO is quite low for wideband speech on the fullband scale since the wideband speech does not contain all the

frequencies of the fullband speech. Furthermore, as stated in the ITU's recommendation P.863.1, the fullband mode POLQA generates rather low MOS-LQO for narrowband and wideband codecs, when compared to subjective scores, MOS-LQS, for the same codecs.

An interesting observation from the result was that the perceived speech quality for the male speaker achieved a lower MOS-LQO than the female speakers. Since the MOS-LQO was similar between the genders for narrowband speech, the difference must be caused by the fullband mode POLQA or/and the VQT. As explained in ITU's recommendation P.863.1 [26], the two POLQA modes contain different optimizations in the sound perceptual model. The optimization for fullband mode may be biased towards female speech since female speech have generally more and higher frequencies than male speech, as mentioned in Section 2.2. Furthermore, fullband POLQA evaluation requires external upsampling, as mentioned in Section 2.6.2, which the VQT does not perform, which may also affect the evaluation results for the wideband speech.

The results of the PESQ evaluation for wideband are seen in Figure 5.23 and Figure 5.24. Since wideband PESQ was developed to only evaluate wideband speech, and not wideband to fullband as POLQA fullband mode, the results of the evaluation were rather high values on the MOS-LQO scale. The male speaker still had a higher MOS-LQO than the female speaker, which can be caused by the different sound levels of the reference signal and the degraded signal, and/or the biased perception model used in PESQ. The effect of enabling VAD decreased the perceived speech quality, which would indicate that some speech was wrongly classified as non-speech.

6.1.4 Overall Performance of G729D and Opus

From the result presented in Chapter 5, Opus performed better than G729D regarding the perceived speech quality with both the POLQA and PESQ evaluation. Therefore, Opus is the preferred speech codec to use for narrowband speech compression.

A bit rate of 8 kbit/s was the lowest bit rate to be used for the speech evaluation with Opus, as recommended from the documentation [16]. For G729D, a 6.4 kbit/s bit rate was used, which means that G729D's result can also be influenced by the difference in bit rates between the codecs. If lower bit rates would be more desired than higher perceived speech quality, G729D could be considered instead of Opus.

According to J. Skoglund and J.-M. Valin's paper [31], the perceived speech quality with Opus would deteriorate below a bit rate of 10 kbit/s. With VAD enabled a similar result was produced, since VAD lowered the perceived quality with a bit rate of 8 kbit/s, which can be seen in e.g. Figure 5.13 and Figure 5.14. Without VAD enabled, a similar MOS-LQO was achieved between the two narrowband bit rates, 8 kbit/s, and 12 kbit/s, which would indicate that the perceived speech quality does not deteriorate as quickly below 10 kbit/s bit rate when VAD is

disabled. Furthermore, similar results as F. Zampognaro et al. paper [32] were achieved, since Opus had better perceived speech quality than G729D.

6.1.5 Evaluation Metrics

One factor that must be considered when reviewing the result is that both PESQ and POLQA are objective models, which estimate what the MOS would have been if a subjective listening test had been performed. Since no subjective listening test was performed for the thesis, there is uncertainty about what the actual MOS-LQO of the audio files is with the different codecs and network scenarios. A MOS-LQS could have been used as a baseline for the perceived speech quality, and the correlation between the MOS-LQS and the two generated MOS-LQO could have been evaluated instead. Since POLQA is recommended and developed for VoIP communication, POLQA's MOS-LQO values are closest to what subjective listening MOS values would have been, as stated in ITU's recommendation [26]. A disadvantage of the PESQ evaluation is that the evaluation metrics do not consider sound level differences between the degraded and reference signal. Therefore, the female speech was classified with a lower MOS-LQO than the male speech for the PESQ evaluation since the gain for the female speech was increased for the degraded signal. If the audio of the female speech was not adjusted, the noise would interfere with the evaluation and the MOS-LQO would decrease for both PESQ and POLQA. Furthermore, the male speech was evaluated higher with dataset 2 even though the dataset did not have a noticeable gain difference. Consequently, the perception model used in PESQ could be biased towards male speech since the female speech with PESQ achieved similar evaluation scores as the female and male speech with POLQA.

From the result of the evaluation of the second dataset, see Section 5.2.2, PESQ achieved higher MOS-LQO than POLQA. Since POLQA should be more accurate than PESQ, according to ITU's recommendation [26], the results would indicate that PESQ overestimated the MOS-LQO for dataset 2.

Lastly, it can be concluded from the results presented in Chapter 5 that PESQ generate a more consistent result for each of the measurements, whilst POLQA had some variation between each measurement. The variation of MOS-LQO was caused by a network switch that was connected to the emulator and used by the VQuad running the POLQA evaluation. This is explained further in the next section, Section 6.1.6.

6.1.6 Network Scenarios and Outliers

The main differences between the different network scenarios were the link speed and the latency. Latency, as mentioned in Section 4.1, is not measured with PESQ and POLQA, therefore it did not affect the results from the evaluation. Consequently, the MOS-LQO for each of the test cases was quite similar for the different network scenarios. The other main difference between the network scenarios was the link speed. As seen from the results, the link speed, nor the loss rate, had

made a significant difference between the measurement.

As observed from the results, some of the network scenarios that should produce similar values of MOS-LQO with POLQA had outliers. The reason for the difference was caused by the network switch connecting to the emulator and the VQuad with the POLQA license. During the testing, it was observed that the scoring interval for POLQA with satellite two-hop, seen in Figure 5.6, for the male speaker, was between 1.15 – 3.19 which produced an average score of 2.8. The MOS-LQO for POLQA could also decrease over time, but after restarting the network emulator the MOS-LQO would jump back up. Another reason for some outliers can be explained by the randomness of the emulated network scenarios. For instance, there could be more frequent packet losses for some emulation runs, but for some other runs, no packets were dropped. Therefore, the differences in the POLQA evaluation is most likely caused by the faulty network switch connected to the emulator.

6.1.7 Voice Activity Detection Effect on the Perceived Speech Quality

The result of enabled voice activity detection for Opus can be seen in figures: 5.5, 5.7, 5.13 and 5.15 for the POLQA evaluation with the two narrowband datasets. For dataset 1, VAD had a positive impact on the MOS-LQO when a 12 kbit/s were used, the MOS-LQO increased with ~0.15. VAD did however not affect the MOS-LQO for dataset 2 with a 12 kbit/s bit rate. The cause of the increased MOS-LQO is that the VAD detected the non-speech periods better, and low-bit comfort noise was sent instead of the encoded noise. Therefore, VAD can lower the amount of transmission noise and improve the MOS-LQO slightly for some files. An observation for the lower bit rate at 8 kbit/s was that the MOS-LQO decreased with 0.3 – 0.5 when VAD was enabled for both datasets, which indicates that VAD could not differentiate speech and non-speech as easily in lower bit rates. Therefore, some speech frames were interpreted as noise and replaced with comfort noise, which lowered the perceived speech quality. The results for wideband speech at 16 kbit/s are seen in Figure 5.21 and Figure 5.23. For POLQA, VAD caused no difference in perceived speech quality when comparing the results seen in e.g., Figure 5.21 and Figure 5.22.

For the PESQ evaluation with the two narrowband datasets, the results of enabled VAD are seen in figures: 5.9, 5.11, 5.17 and 5.19. For the 12 kbit/s bit rate, a small increase of 0.1 MOS-LQO can be seen for the female speaker and no effect for the male speaker when VAD was enabled for dataset 1. A small increase > 0.1 MOS-LQO could also be seen for dataset 2 for both the male and female speakers with a 12 kbit/s bit rate. As mentioned in the paragraph above, the increased MOS-LQO is caused by VAD replacing non-speech frames with low-bit comfort noise. A decrease of 0.3 – 0.6 MOS-LQO was observed for dataset 1 with an 8 kbit/s bit rate and about an 0.5 MOS-LQO decrease for dataset 2, which is caused by VAD wrongly classifying some speech frame as non-speech and replaces them with the low-bit comfort noise. The wideband PESQ evaluation with and without

VAD enabled can be seen in Figure 5.23 and Figure 5.24. For PESQ, the perceived speech quality was lowered when VAD was enabled for wideband speech, about 0.05 for the male speech and about 0.16 for the female speech. The lower MOS-LQO for the PESQ evaluation are considered to be less accurate when compared to POLQA since PESQ is not adapted for VoIP.

Consequently, VAD can be used to lower the total amount of bits needed for the data transmission without affecting the perceived speech quality negatively when a 12 kbit/s or higher bit rate is used. For lower bit rates, VAD affected the perceived speech quality negatively and therefore should VAD be avoided for Opus with a bit rate of 8 kbit/s.

6.2 Choice of Method

In the following section, the choice of method is evaluated and criticized. The replicability, reliability, and validity of the implementation and evaluation are discussed. Lastly, the sources used for the thesis are motivated.

VoIP Implementation

Low-level APIs of the PJSIP library were used to gain a better understanding of the library and how codecs are integrated. A high-level API, PJSUA, could have been used for the implementation. However, with a high-level API, it would have been more difficult to understand how the low-level APIs calls are used and how to integrate the codecs properly. The integration of the codecs into PJSIP took longer time than expected which resulted in time deficiency and MELPe was not implemented in time for the evaluation. As mentioned in Section 4.4, another SIP library could have been used instead of PJSIP. Linphone was investigated, however, the documentation was not as well written as PJSIP's and did not contain defined instructions for third-party codec integration. Therefore, PJSIP was the better choice to use for the thesis as a SIP library.

Unencrypted Data

For the implemented VoIP application, the streamed data was not encrypted. When data is streamed over the internet, all data is usually encrypted to keep the information safe and protected from manipulation. The main type of encryption used is AES, a type of lossless encryption. Consequently, the evaluation of the codecs with encryption would only increase the latency, which would not affect the MOS-LQO. Hence, it was decided to keep the data unencrypted for the codec evaluation.

Test System

When it comes to the wideband speech evaluation with POLQA and the VQT, a few things still do not add up. According to the POLQA documentation [21], the

16 kHz degraded signal should be upsampled to 48 kHz such that it can be compared to the reference signal. However, according to GL Communication, the degraded signal is not upsampled for the POLQA evaluation. This means that the reference signal contains three times more sample points than the degraded signal in the evaluation. GL Communication was reached to provide some explanation, but the received response did not give any further explanation to their POLQA evaluation. It is highly likely that the degraded 16 kHz signal is interpolated between its sample points, or that every third sample point of the reference signal is used for the evaluation. Moreover, the 16 kHz sampled signal contained fewer frequencies than the 48 kHz sampled signal. As mentioned in 2.2, the human voice contains frequencies over 8 kHz, which are filtered out for the 16 kHz sampled signal but not for the 48 kHz sampled audio. Consequently, the resulted MOS-LQO from the evaluation will always be lower for the wideband speech when compared with a fullband signal.

Audio Datasets

The audio datasets could have been chosen more wisely. The audio files used were provided through the test equipment, the VQuad, and therefore it was assumed that the files would be good to use for the evaluation. However, during the evaluation with dataset 1, the sound level was different between the two speech files, which affected the MOS-LQO negatively for the female speech. The second dataset was better suited for the evaluation since the gain was the same for both the male and female speech files, which can be seen in the results of the PESQ evaluation. The provided audio from the VQuad did not contain dataset 2, the gibberish speech, for higher sample rates than 16 kHz. Since POLQA needed a reference signal with a 48 kHz sample rate for wideband speech, these speech files could not be tested for wideband speech. Instead, dataset 1 was used to evaluate the wideband speech, since the VQuad contained the correct sample rates for these files.

Network Conditions

The network scenarios used in the emulator were created to mimic the actual network conditions for 2G, 3G, 4G, satellite two-hop, and LAN, and therefore they were used for the evaluation of the codecs' perceived speech quality. As explained previously, the different scenarios did not affect the MOS-LQO for the codecs used, due to the similar network settings used and latency not being considered in the evaluation. The settings for the network scenarios could have been tweaked to give more diverse results e.g., increase the packet loss for the satellite network.

6.2.1 Replicability, Reliability, and Validity

The work presented in the thesis can be replicated with PJSIP and integrating G729D and Opus into the library. Both G729D and Opus are open-source codecs that can be downloaded from ITU respectively Opus website. To evaluate the

codecs with POLQA and PESQ wideband licenses must be bought. PESQ for narrowband is open-sourced and can be downloaded from ITU. The audio files used for the test and evaluation can be obtained from the POLQA and PESQ licenses to evaluate the codecs with the same audio files used in the thesis.

The results obtained from the evaluation can be reliable in such a sense that the performance measurements, PESQ and POLQA, are developed for evaluation of, among other things, codecs. The same results can be produced if the same audio files and network scenarios are used.

To validate the obtained results, more speech files could have been used. In total eight different speech files were used, four male and four female speakers. This gave a quite narrow spectrum of different voices, and as can be seen from the results of dataset 1, there were some differences between the perceived quality between the male and female speakers. A larger set of English speech files of both male and female voices could be used to better assess if the difference in perceived speech quality is present for more speakers for different bandwidths. To better validate the achieved MOS-LQO from the evaluation, the speech files should have been labeled with a subjective listening MOS, MOS-LQS. A MOS-LQS would have been a better representation of the perceived speech quality, which could have been compared to the MOS-LQO from PESQ and POLQA to achieve a better evaluation of G729D and Opus.

6.2.2 Source Criticism

In the background chapter, Chapter 2, the sources used were primarily books and ITU recommendations. Most of the books have been used as course literature, which indicates that the information is trustworthy. Besides books, ITU recommendations are highly used in the background since ITU sets the standards for telecommunication, and most papers presented as related work uses their recommendations.

For the related work chapter, Chapter 3, the intention was to pick sources that were either from high-ranked conferences/journals or well cited. Since the subject of perceived speech quality and compression has been well researched, some sources that are relevant for the thesis have been published in older editions of high-ranked conferences. These editions usually do not have the same ranking as today's edition, which decreases the credibility of the source. However, the sources chosen were picked to give the reader an insight of what has been researched regarding codecs and perceived speech quality. An example of these sources is P. Souček and J. Holub's paper [33], which is outdated since POLQA have been updated after the paper was published in 2012. However, the paper was chosen to give the reader a wider context of the QoE metric. Both codecs and QoE metrics are developed with a certain amount of speech files and a certain diversity of languages. Since languages have different vocal characteristics, the languages used for the perception model of the codec or metric perform better than other languages.

6.3 Wider Context to the Presented Work

Codecs should be used to lower the bits needed to transmit speech over the internet. With decreased number of bits needed to send the data, more people would have access to information even in underdeveloped areas or areas with limited internet access/low bandwidths.

Furthermore, voice transmission is no longer confined to narrowband, which means that better perceived speech quality is possible to achieve with higher bandwidths. Codecs can be used, as mentioned, for VoIP but also for video over IP. The digital meeting space is continuing to evolve, where better sound and video can be achieved with lower bandwidths[7] with the help of codecs.

Most speech is transmitted with codecs on wideband or higher bandwidths to increase the perceived speech quality. However, most areas with high security still use narrowband for speech transmission, like NATO and US military which uses MELPe [15], since it is more secure. Therefore, the codecs used in narrowband must be developed and improved to increase the quality for lower bandwidths such that old codecs can be replaced. As seen in the results, Opus could be used instead of G729D for lower bandwidths, since the perceived speech quality is better with Opus.

Codecs are important for other speech-related research areas like speech recognition and emotion identification. Without good perceived speech quality will speech intelligibility decrease which would impact what information that can be retrieved from the speech. For instance, a type of speech recognition to recognize Covid-19 from speech [40] has been developed to help identify the disease, and without good speech intelligibility would the classification be impossible to achieve.

[7] https://www.businessinsider.co.za/heres-why-your-video-conference-app-keeps-acting-u accessed 2022-05-12

7

Conclusions and Future Work

The following chapter summarizes and concludes the work presented in the thesis based on the research questions, as seen in Section 1.3. Lastly, some future work of the thesis is introduced, which would wider the evaluation of audio codecs and their perceived speech quality.

7.1 Conclusions

The research questions are answered as follows:

What QoE metrics should be used to evaluate the perceived speech quality of G729D and Opus?

When evaluating the perceived speech quality for audio codecs, a listening only test should be performed. The recommended evaluation metrics are subjective listening tests since a MOS-LQS directly from subjects is the most accurate. However, since these tests are quite expensive, objective tests can be performed instead. The most used and recommended objective speech quality test is POLQA, since it is developed for VoIP transmissions and have better correlation with subjective tests than its predecessor, PESQ. PESQ can be used for evaluating perceived speech quality, but from the result seen in Chapter 5 and explained in Section 2.6.2 are the results not as reliable as the MOS-LQO from the POLQA evaluation.

For wideband speech evaluation, both the PESQ evaluation and POLQA evaluation have some negative aspects. For POLQA, the degraded signal is compared with a fullband reference signal, sampled at 48 kHz, which means that the resulted MOS-LQO is on the fullband scale and some of the highest frequencies of

the 48 kHz signal were filtered out for the 16 kHz signal. Therefore, the MOS-LQO will be low, since wideband audio will always produce lower quality than fullband audio on the fullband scale. For PESQ, the resulted MOS-LQO is calculated for wideband speech quality, which is why the MOS-LQO is higher than for POLQA. However, PESQ is not as well adapted for VoIP which can affect the results. Lastly, for wideband speech, both metrics can be used, but they cannot be compared due to the different scales used.

How does Opus perform in comparison to G729D in different network scenarios regarding the perceived speech quality with narrowband bandwidth

Since latency nor the link speed had any effect on the results of the QoE evaluation, the results between the different network scenarios are negligible. Even so, from the result presented in Chapter 5, Opus outperformed G729D with higher MOS-LQO for both PESQ and POLQA for both datasets. However, in a few instances, G729D performed equally or better than Opus. With VAD enabled for 8 kbit/s bit rate, the achieved MOS-LQO was quite similar between Opus and G729D for dataset 2, and for dataset 1 did G729D achieve a higher MOS-LQO than Opus with VAD enabled.

How is the perceived speech quality affected when an 8 kbit/s and a 12 kbit/s bit rate on narrowband is used with Opus for different network scenarios?

For narrowband speech with an 8 kbit/s and a 12 kbit/s, without VAD, the perceived speech quality was similar for both POLQA, seen in figures: 5.6, 5.8, 5.14, 5.16, and PESQ, seen in figures: 5.10, 5.12, 5.18, 5.20. As mentioned previously, the different network scenarios did not generate any difference of MOS-LQO. The results would indicate that Opus with an 8 kbit/s bit rate can be used instead of a 12 kbit/s bit rate without loss of perceived speech quality if VAD is disabled.

How is the perceived speech quality affected when a 16 kbit/s wideband is used for Opus for different network scenarios?

As concluded, the different network scenarios did not affect the MOS-LQO evaluation. For the POLQA evaluation, the perceived speech quality was low since the perceived speech quality was evaluated with a degraded wideband speech signal to a fullband reference signal. However, an interesting observation with the wideband POLQA evaluation was that the perceived speech quality for the male speaker was lower than the female speaker. This would indicate that the perception model used in POLQA is biased towards female speech e.g., speech with more and higher frequencies. The perceived speech quality with PESQ generated a more reasonable MOS-LQO than POLQA since the degraded wideband speech signal was compared with a wideband reference signal instead of a fullband reference signal. The female speech was rated with a lower MOS-LQO with PESQ, similar to the result of the PESQ evaluation seen for narrowband speech, which was caused by the sound level difference and that the perception model used for PESQ seems to be optimized for male speech which contains lower frequencies.

Does voice activity detection (VAD) have any impact on the perceived speech quality

for Opus with different bit rates?

The most significant difference of the MOS-LQO with VAD enabled was with the lowest tested bit rate, 8 kbit/s. The MOS-LQO decreased with VAD enabled, which would indicate that it was harder for VAD to determine what frames contain speech and which frames contain non-speech, and therefore was some speech frames replaced with comfort noise. For a 12 kbit/s narrowband speech, VAD increased the perceived speech quality slightly. These results would indicate that an increased bit rate improved the classification for the VAD for narrowband speech.Furthermore, when VAD was enabled for the POLQA wideband speech evaluation with a 16 kbit/s bit rate, the MOS-LQO was not affected, which means that the perceived speech quality is not affected by VAD for higher bit rates. For the PESQ evaluation, a decreased MOS-LQO was observed. The cause of the lower MOS-LQO, as stated previously, is that PESQ is not adapted for VoIP communication evaluation and therefore is the result assumed to be less accurate.

7.2 Future Work

Since Opus is recommended to be used with a bit rate interval between 8 kbit/s and 12 kbit/s for narrowband speech, the bit rate used for the evaluation was within this interval. However, Opus can compress audio with lower bit rates, minimum at 6 kbit/s. It would be interesting to evaluate how the perceived speech quality would be affected when lowering the bit rate for Opus to match G729D's bit rate of 6.4 kbit/s. For wideband speech, only the 16 kbit/s bit rate was evaluated with Opus, which is the lower limit recommended for wideband speech. It would have been interesting to evaluate the upper limit to wideband speech, 20 kbit/s, to compare the difference of MOS-LQO for two different bit rates for wideband speech.

Furthermore, a codec that operates at even lower bit rates than G729D and Opus is MELPe. It would be interesting to extend the work presented in the thesis to evaluate how MELPe would perform compared with G729D and Opus. Another codec that would have been interesting to evaluate its perceived speech quality in lower bit rates is 3GPP's codec EVS (enhanced voice service). EVS, like Opus, was developed to operate a wide range of bandwidths. However, the codec is not as established as Opus and was not available to use for evaluation for the thesis and therefore would be possible to evaluate as future work.

For the evaluation of the codecs two intrusive metrics were used, PESQ and POLQA. For real-world scenarios, the reference signal will most likely not be possible to use for evaluation. The current research is mostly regarding non-intrusive data-driven models to predict the QoE score. Therefore, to extend the thesis the QoE can be evaluated with some of these data-driven models, which would better reflect how to measure the perceived speech quality for real-world scenarios. Furthermore, the MOS-LQO of the data-driven models could be compared to MOS-LQO generated by PESQ and POLQA to investigate if similar results or better results can be achieved by the data-driven models.

An interesting factor that could be further investigated is how much the energy consumption changes when transmitting speech with different bit rates and bandwidths. Since more data is transmitted for higher bit rates, it should yield higher energy consumption. Furthermore, it would be interesting to evaluate the energy consumption between the codecs when using the same bit rate, 6.4 kbit/s, and for Opus 8 kbit/s and 12 kbit/s for narrowband and 16 kbit/s for wideband. For VoIP communication, the demand for high QoE is increasing as higher bandwidths are used. Therefore, it is important to further investigate energy consumption and perceived speech quality such that the end-users can be satisfied with the QoE without large energy losses e.g., battery drainage in telephones contra perceived speech quality during phone calls.

As can be seen in figures: 5.5, 5.13, 5.9 and 5.17, VAD for Opus for 8 kbit/s lowered the perceived speech quality. In future work, the perceived speech quality with G729D with VAD, also known as G729F, could be evaluated. Lastly, other VAD algorithms could be evaluated for low-bit-rate codecs to investigate if other VAD algorithms can further decrease the total number of bits needed without lowering the perceived speech quality.

Bibliography

[1] K. Sayood, *Introduction to Data Compression*. Elsevier Science & Technology, third ed., 2005.

[2] H. P. Singh, S. Singh, J. Singh, and S. Khan, "Voip: State of art for global connectivity—a critical review," *Journal of Network and Computer Applications*, vol. 37, pp. 365–379, 2014.

[3] ITU-T Recommendation P.800.1, "Mean opinion score (mos) terminology," *International Telecommunication Union*, 2016.

[4] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, pp. 36–41, March 2010.

[5] F. A. Everest and K. C. Pohlmann, *Master handbook of acoustics*. McGraw-Hill Education LLC., 2015.

[6] D. Salomon and G. Motta, *Handbook of data compression*. Springer, 2010.

[7] S. C. Levy, D. J. Freed, M. Nilsson, B. C. J. Moore, and S. Puria, "Extended high-frequency bandwidth improves speech reception in the presence of spatially separated masking speech," *Ear and Hearing*, vol. 36, p. e214–e224, 2015.

[8] E. Pépiot, "Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers," in *Speech Prosody 7*, (Dublin, Ireland), pp. 305–309, May 2014.

[9] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, 2014.

[10] B. B. Monson, A. J. Lotto, and B. H. Story, "Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1754–1764, 2012.

[11] R. I.-T. P.10/G.100, "Vocabulary for performance, quality of service and quality of experience," *International Telecommunication Union*, 2017.

[12] M. Goncalves, *Voice Over IP Networks.* McGraw-Hill, 1999.

[13] A. Kumar, "Study of different types coders for gsm," in *International Journal of Engineering and Technical Research (IJETR)*, vol. 2, 2014.

[14] ITU-T Recommendation G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (cs-acelp)," *International Telecommunication Union*, 2012.

[15] Compandent, "Stanag-4591 melpe - enhanced mixed-excitation linear predictive vocoder software," (accessed: 25.01.2022).

[16] Internet Engineering Task Force (IETF), "Definition of the opus audio codec," 2012. (accessed: 25.01.2022).

[17] A. Raake, *Speech Quality of VoIP: Assessment and Prediction.* Wiley, 2006.

[18] B. A. Forouzan, *TCP/IP protocol suite.* McGraw-Hill Forouzan networking series, McGraw-Hill, 2010.

[19] ITU-T Recommendation P.800.2, "Mean opinion score interpretation and reporting," *International Telecommunication Union*, 2016.

[20] ITU-R Recommendation BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," *International Telecommunication Union*, 2005.

[21] ITU-T Recommendation P.863, "Perceptual objective listening quality prediction," *International Telecommunication Union*, 2018.

[22] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *Proc. Interspeech 2020*, pp. 4631–4635, 2020.

[23] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7125–7129, 2019.

[24] A. Hines, E. Gillen, and N. Harte, "Measuring and monitoring speech quality for voice over IP with POLQA, viSQOL and p.563," in *Proc. Interspeech 2015*, pp. 438–442, 2015.

[25] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.," *International Telecommunication Union*, 2001.

[26] ITU-T Recommendation P.863.1, "Application guide for recommendation itu-t p.863," *International Telecommunication Union*, 2019.

[27] ITU-T Recommendation P.862.1, "Mapping function for transforming p.862 raw result scores to mos-lqo," *International Telecommunication Union*, 2003.

[28] ITU-T Recommendation P.862.2, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union*, 2007.

[29] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *International Telecommunication Union*, 2004.

[30] ITU-T Recommendation G.107, "The e-model: a computational model for use in transmission planning.," *International Telecommunication Union*, 2015.

[31] J. Skoglund and J.-M. Valin, "Improving Opus Low Bit Rate Quality with Neural Speech Synthesis," in *Proc. Interspeech 2020*, pp. 2847–2851, 2020.

[32] F. Zampognaro, R. Aurigemma, and W. Munarini, "Voip codec assessment and performance evaluation in satellite-based scenarios," in *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, pp. 01–06, 2021.

[33] P. Souček and J. Holub, "Evaluation of itu-t p.863 polqa in chinese environment," in *2012 IEEE 1st International Symposium on Wireless Systems (IDAACS-SWS)*, pp. 124–126, 2012.

[34] J. Kaiser and T. Bořil, "Impact of the gsm amr codec on automatic vowel formant measurement in praat and voicesauce," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pp. 1–4, 2018.

[35] T. Volk, C. Keimel, M. Moosmeier, and K. Diepold, "Crowdsourcing vs. laboratory experiments – qoe evaluation of binaural playback in a teleconference scenario," *Computer Networks*, vol. 90, pp. 99–109, 2015. Crowdsourcing.

[36] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[37] N. Nessler, M. Cernak, P. Prandoni, and P. Mainar, "Non-Intrusive Speech Quality Assessment with Transfer Learning and Subject-Specific Scaling," in *Proc. Interspeech 2021*, pp. 2406–2410, 2021.

[38] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *journal of the audio engineering society*, vol. 61, pp. 366–384, june 2013.

[39] F. Köster and S. Möller, "Analyzing the Relation Between Overall Quality and the Quality of Individual Phases in a Telephone Conversation," in *Proc. Interspeech 2016*, pp. 2493–2497, 2016.

[40] I. Södergren, M. P. Nodeh, P. C. Chhipa, K. Nikolaidou, and G. Kovács, "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines," in *Proc. Interspeech 2021*, pp. 916–920, 2021.