Doctoral Thesis

A Study of the Effect of Noise and Network Distortions in VoIP Speech Signals

by

Elhard James Kumalija

September 2023

Graduate School of Applied Informatics

University of Hyogo

Abstract

Voice over IP (VoIP) is used in a broad number of services. In VoIP applications, such as Interactive Voice Response and VoIP-phone conversation transcription, speech signals are degraded not only by environmental noise but also by transmission network errors, distortions induced by encoding and decoding algorithms.

The effect of noise and network distortions on VoIP speech application depends on the nature of application. For example, in public broadcasting, we are concerned with the speech quality received by announcement listeners. Therefore, there is a need for a continuous automatic quality evaluation of the transmitted audio. Speech quality monitoring in VoIP systems enables autonomous system adaptation to network changes or environmental noise level changes. While, in VoIP-phone transcription or Interaction voice response, the main concern is the accuracy of the speech recognition system. Furthermore, there are diverse IP audio transmitters and receivers, from high-performance computers and mobile phones to low-memory and low-computing-capacity embedded systems.

This study intends to understand the effect of noise-network distortions on automatic speech recognition and evaluation of speech quality in VoIP application. Furthermore, this work proposes the method to develop highly robust automatic speech recognition systems and speech quality prediction models.

Firstly, a comparative analysis of a speech-to-text system trained on clean speech against one trained on integrated noise-network distorted speech is presented. Training an ASR model on noise-network distorted speech dataset improves its ability to cope with distorted speech without loss of accuracy. Although the performance of an ASR model trained on clean speech depends on noise type, this is not the case when noise is further distorted by network transmission.

The model trained on noise-network distorted speech exhibited a 60% improvement rate in the word error rate (WER), word match rate (MER), and word information lost (WIL) over the model trained on clean speech. Furthermore, the ASR model trained with noise-network distorted speech could tolerate a jitter of less than 20% and a packet loss of less than 15%, without a decrease in performance. However, WER, MER, and WIL increased in proportion to the jitter and packet loss as they exceeded 20% and 15%, respectively. Additionally, the model trained on noise-network distorted speech exhibited low accuracy loss compared to that trained on clean speech. The ASR model trained on noise-network distorted speech can also tolerate signal-to-noise (SNR) values of 5 dB and above, without the loss of performance, independent of noise type.

Secondly, MiniatureVQNet a single-ended speech quality evaluation method for VoIP audio applications based on a lightweight deep neural network (DNN) model is proposed. This is followed by the performance analysis of the MiniatureVQNet model trained on clean speech compared to the MiniatureVQNet model trained on integrated noise-network distorted speech dataset.

The proposed model can predict the audio quality independent of the source of degradation, whether noise or network, and is light enough to run in embedded systems. Two variations of the proposed MiniatureVQNet model were evaluated: a MiniatureVQNet model trained on a dataset that contains environmental noise only, referred to as MiniatureVQNet–Noise, and a second model trained on both noise and network distortions, referred to as MiniatureVQNet–Noise–Network.

The proposed MiniatureVQNet model outperforms the traditional P.563 method in terms of accuracy on all tested network conditions and environmental noise parameters. The mean squared error (MSE) of the single-end speech quality prediction methods ITU-T P.563, MiniatureVQNet–Noise, and MiniatureVQNet–Noise–Network compared to the PESQ (end-to-end method) score for was 2.19, 0.34, and 0.21, respectively.

The performance of both the MiniatureVQNet–Noise–Network and MiniatureVQNet–Noise model depends on the noise type for an SNR greater than 0 dB and less than 10 dB. In addition, training on a noise–network–distorted speech dataset improves the model prediction accuracy in all VoIP environment distortions compared to training the model on a noise-only dataset.

Acknowledgements

I thank the almighty God for his everyday grace and blessings which enabled me to meet Professor Yukikazu Nakamoto, to do, and to accomplish this research successfully.

I am very much grateful to my research advisor Prof. Yukikazu Nakamoto for his tireless guidance and the support he accorded me during all the years of my research. There are moments when I was in total despair, but his advice and encouragement restored hope and rekindled the desire to continue.

I would like to thank Prof. Hiroaki Ohshima of the University of Hyogo for his for constructive suggestions, guidance and time devoted in this research work. My thanks are extended to Professor Ryo Haraguchi, his interest and curiosity on my research work was very helpful in shaping the final presentation of my research work. Special thanks go to Prof. Jun Kurihara for his constructive advice on my research work.

It is my pleasure to thank the following individuals who supported me in different ways in this study; Prof. Yi Sun of the Kobe Institute of Computing, Miss Suzan Kessy of the Nara Institute of Technology, Hiroaki Kubota of TOA Corporation, and Godwin Tunze. I also thank Mr. Koji Doi and all University of Hyogo staff members for the encouragement and support in the time of need, and all who in one way or the other have contributed in this study.

Unique appreciation should go to my family for their prayers, encouragement, support, understanding and patience during the whole period of this study. Special thanks to my youngest daughter Gianna Kumalija who were born during this study. There are times she missed the father's care because of the time dedicated to this study.

Contents

List of Figures

List of Tables

List of Equations

CHAPTER 1 INTRODUCTION

IP network is a communication network that uses Internet Protocol (IP) to send and receive messages between one or more computers. As one of the most commonly used global networks, an IP network is implemented in Internet networks, local area networks (LAN) and enterprise networks. Widespread availability of IP network has driven integration of different types of communication into IP network, including Voice over Internet Protocol (VoIP).

Today, Voice over Internet Protocol (VoIP) is the most applied method for voice communication transmission because of the wide availability of broadband IP network. VoIP services can be offered at low price compared to Public switched telephone networks (PSTN). An increase in IP network bandwidth, decrease in cost, and increase in the accessibility of the IP network has enabled the introduction of more IP audio services beyond merely voice call. Internet radios and music streaming applications are examples of such services.

VoIP is a common component of applications such as social networking services, internet radio, and multimedia streaming applications. Compared to PSTN, VoIP services can be offered on a wide variety of devices such as IP phones, smartphone, and personal computers. Thus, there are many services that can be offered by VoIP.

However, the IP network is a best effort network, the packet delivery is not guaranteed, but depends on the availability of resource to transfer the packet at the time. Hence, in additional to the effect of environmental noise on the captured audio signal, the quality of audio signal in VoIP is affected by network condition.

1.1 VoIP Application Ecosystem

The audio signal transmitted on VoIP application is affected by environmental noise and network distortions as Figure 1 shows. The voice captured for transmission in VoIP system is not only the desired speakers' voice but also some background noises e.g. engine noise captured during hands free in-vehicle voice call.

The IP network is a best effort network. Hence, there are network transmission errors when speech audio is transmitted through IP networks. Codecs and Compression, Packet loss,

latency, jitter, and Out-of-Order packets are the common IP network transmission errors on VoIP speech signal.

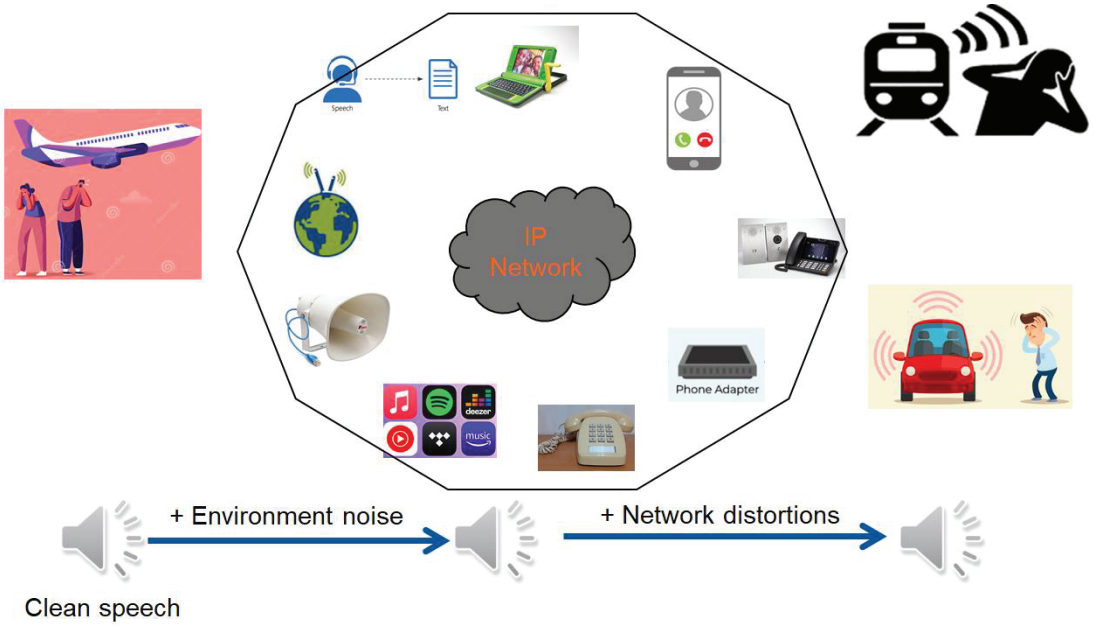

Figure 1 VoIP Application Ecosystem

The effect of audio quality degradation depends much on the nature of the application. For example, in two way communication between human beings, the degradation can be observed and communication participants can decide mitigation measures, such as rescheduling the meeting to a less network congested time. However, in one-way communication, which is a common form of VoIP streaming in public addressing systems, and internet radio, sender cannot get feedback from the receiver, and in many cases there are many VoIP stream recipients for one sender. In this case, many content delivery networks (CDN) detects the network quality of connected node and transcodes the stream with bitrate appropriate for the node. Alternatively, the CDN transcodes a set of streams for different network bandwidths, and receiving clients can choose the stream based on their network conditions.

Therefore, there is a need to automatically monitor the audio quality in VoIP applications. Quality parameters obtained from automatic monitoring of audio quality in VoIP can be

used to re-route the audio packets, changing audio codec parameters, changing parameters of noise reduction systems, and etc.

On the other hand, the quality degraded audio in VoIP is an input to other systems, such as automatic speech recognition (ASR), interactive voice response (IVR), and voice transcription systems. When the audio speech signal input is degraded by Environmental noise and network distortion on VoIP audio may lead to inaccurate results or performance decrease in different operating environment, considering the wide variety of applications and devices in VoIP systems. In this case, the systems need to mitigate the effect of audio degradation caused by network and environmental noise distortions in order to increase their ability to perform well in noise-network distorted speech environment without loss of accuracy. Moreover, compared to PSTN, VoIP services can be offered on a wide variety of devices such as IP phones, smartphone, and personal computers. These devices vary a lot in terms of computing power, memory, usage, and usage environment. Although, this is an advantage of VoIP, on the contrary, this also introduces challenges in support and interoperability of the devices.

1.2 Tackling Environmental Noise and Network Distortion in VoIP

Tackling environmental noise and network distortion in VoIP communication is essential to ensure high-quality audio and a seamless user experience. There are numerous studies on different strategies to address environmental noise and network distortion in VoIP as shown in Figure 2.

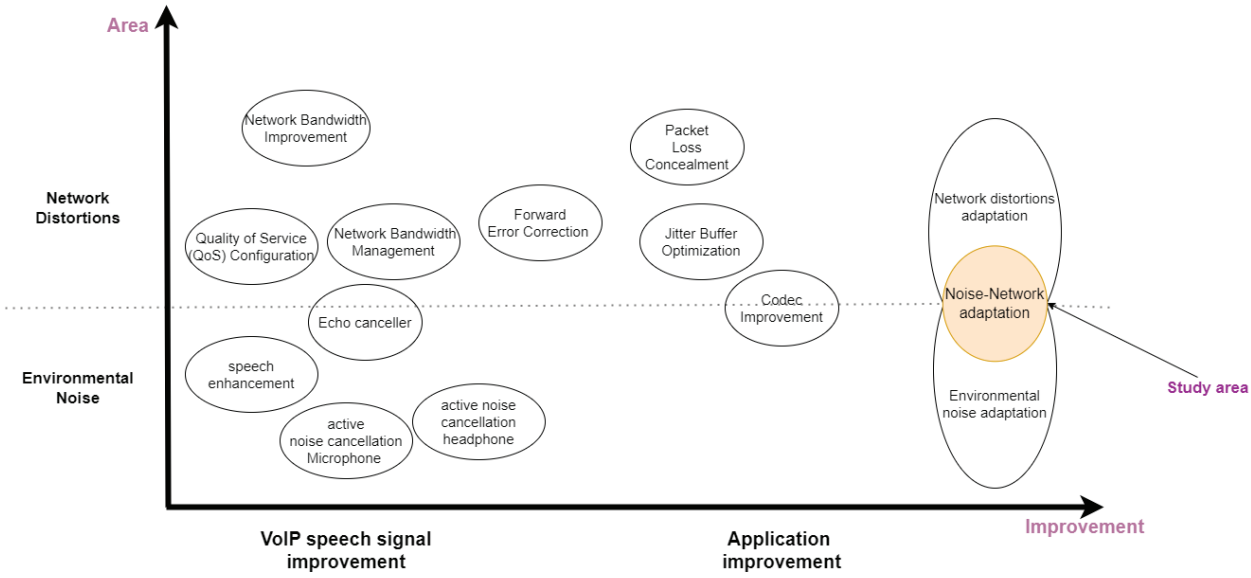

Figure 2 Tackling Environmental Noise and Network Distortion in VoIP

1.2.1 Strategies to Address Environmental Noise in VoIP

Environmental noise in VoIP refers to unwanted sounds or disturbances from the surrounding environment that can negatively affect the quality of voice communication during VoIP calls. This noise can interfere with the clarity of the speaker's voice, making it difficult for participants to understand each other. To address environmental noise different algorithms are used to remove noise from desired speakers' voice during capturing (using noise cancelling microphones) or after capturing using speech enhancement algorithms.

1.2.1.1 Noise-Canceling Microphones and Headphones

Noise cancelling microphones and headphones employ active noise cancellation (ANC) techniques to attenuate environmental noise (Liu et al., 2023), (Molesworth et al., 2013). ANC is more effective at reducing low frequency sounds (Ang et al., 2017).

1.2.1.2 Speech Enhancement

The goal of speech enhancement is to make noisy degraded speech signals more clear, natural, and understandable by reducing or eliminating various types of interference such as: background Noise, reverberation, channel distortions, cross-talk and codec artifacts. Removing background noise using audio signal processing techniques include.

Classical signal processing methods: Wiener Suppression rule (Wiener, 1949), and Spectral Subtraction (Loizou, 2013).

Statistical based methods: Maximum likelihood, Short-term minimum mean-square estimator (MMSE), Short-term Log-MMSE, Joint Maximum a-Posteriori Spectral Amplitude and Phase (JMAP SAP) (Loizou, 2013).

Machine learning methods: There are different developed DNN models for speech enhancement using different network architecture. In Xu's work, the regression-based speech enhancement framework (Xu et al., 2014) was proposed, later DNN speech enhancement self-optimized by reinforcement learning (Koizumi et al., 2017) was studied. Moreover, other works are based on different network architecture like Long short-term memory (Weninger et al., 2015) and Generative Adversarial Network (Pascual et al., 2017).

1.2.2 VoIP Measure to Avoid Network Distortions

Currently there are different measures implemented to tackle the effect of IP network on VoIP call quality. The objective of these measures relies on improving the capacity or efficiency utilization of network carrying VoIP packets to avoid network distortions and ensure a smoother, higher-quality VoIP experience for users.

1.2.2.1 Quality of Service (QoS) Configuration

Implement QoS settings to prioritize VoIP traffic over other types of data on IP networks. This ensures that VoIP packets receive preferential treatment and are delivered with minimal delay and jitter.

1.2.2.2 Network Bandwidth Improvement and Management

Network bandwidth improvement and management ensures that there is sufficient available bandwidth to accommodate VoIP traffic without causing congestion or slowdowns. Regularly monitor network usage to prevent overloading.

1.2.2.3 Packet Loss Mitigation

Utilization of technologies like Forward Error Correction (FEC) and Packet Loss Concealment (PLC) to recover lost packets and maintain call quality, even in the presence of network packet loss.

1.2.2.4 Jitter Buffer Optimization

Adjust the jitter buffer size appropriately to account for network jitter. A larger buffer can help smooth out variations in packet arrival times and reduce the impact of jitter on call quality. Jitter buffer management algorithms.

1.2.2.5 Codec Improvement

Choose VoIP codecs that strike a balance between bandwidth efficiency and call quality. Some codecs are more resilient to packet loss and network variations. For Example Support for 8 kHz (Narrowband) to 48 kHz (Fullband), dynamically adjustable bitrate, audio bandwidth, and frame size, and packet loss concealment (PLC).

1.3 Tackling Environmental Noise and Network Distortion in VoIP Applications

Noise and network distortions are not completely eradicated by speech enhancement or network improvement. Remnants of noise and network distortions do reach the final receiver application.

Humans can tolerate noise and network distortions and can easily separate the noise and network distortions effects from the transmitted VoIP speech. However, for VoIP automated services such as call transcription, systems does not have human's ability and their performance degrades on such conditions.

In this thesis, we focus on studying the effect of audio quality degradation caused by environmental noise and network distortions in VoIP application, with the goal of developing mitigation measures that can be applied in a wide array of VoIP audio devices. Our work concentrated on single-ended method for speech quality monitoring and Text-to-speech applications as representative of then many other VoIP applications. There are many studies on the effect measure of environment noise in VoIP applications, and also there are a lot of studies on the effect of network quality in VoIP. However, these effects are not mutually exclusive, but little work has been done on studying the mutual effect of environmental noise and network distortion which is the main focus of this work.

1.4 Research Objective

The main objective of this research work is to analyze the audio quality degradation effect of integrated noise and network distortion in VoIP applications and develop mitigation measures applicable for a wide variety of VoIP equipment.

1.5 Specific Objectives

The specific objectives of this research are:

1. Analysis of the effect of audio degradation caused by the combination of environmental noise and network distortions on VoIP based text-to-speech application.

2. Mitigating the effect of audio degradation caused by the combination of environmental noise and network distortion on VoIP based text-to-speech application.

3. Analysis of the effect of audio degradation caused by the combination of environmental noise and network distortions on VoIP based speech quality monitoring application.

4. Mitigating the effect of audio degradation caused by the combination of environmental noise and network distortion on VoIP based speech quality monitoring application.

1.6 Significance of the Research

First, this thesis presents the systematic analysis of the effect of audio degradation caused by the combination of environmental noise and network distortions on VoIP based application. With the increase in computing mobility, VoIP services are offered in different computing device and environments. One on the go, a mobile phone network changes, users are subjected to different environmental noise. This analysis is important for designing and planning VoIP-based ASR systems, such as Interactive Voice Response, as well as VoIP-phone conversations transcription. Furthermore, the effect of noise-network distortion is analyzed on single-ended speech monitoring methods.

Secondly, this work discusses the potential performance optimization of the existing ASR models pre-trained on clean speech datasets by re-training the models using integrated noise-network distorted speech, and its limitations. Using transfer learning, the

performance of the existing ASR models can be optimized to robustly handle noise-network distorted speech. This method can be replicated in a different problem domain, related to noise and network distortions problems.

This work has resulted in generation of noise-network distorted speech dataset which is a contribution to the academic community. The dataset can enable further studies of the effect of noise and network condition in VoIP applications.

Finally, this thesis proposes MiniatureVQNet, a single-ended speech quality evaluation method for VoIP audio applications based on a lightweight deep neural network (DNN) model. The proposed model can predict the audio quality independent of the source of degradation, whether noise or network, and is light enough to run in embedded systems.

We examined the proposed model performance on different VoIP speech degradation factors, including network distortion and acoustic/environmental noise. Furthermore, different post-training optimization methods were considered to improve the performance in low-resource computing environments such as embedded systems.

1.7 Thesis Outline

This thesis contains three main sections. First section includes three chapters, chapter 1 to chapter 3, which lays the foundation for the study and the relation between this study and other research works related to mitigating the effect of environmental noise and network distortions VoIP. The second part is the main work of this study which includes chapter 4 to chapter 7. In this part analysis of the effect of environmental noise and network distortion on VoIP is done and mitigation measures are proposed. Chapter 8 discusses the results of this work and proposes future direction for research.

Chapter 1 discusses the background of this work. In this chapter the introduction VOIP systems is presented, and the automated services in VoIP applications are discussed. The importance of automated service such as speech quality monitoring and automatic speech recognition is highlighted. Furthermore, the effect of environmental noise and network distortion on VoIP speech signal in relation to the automated VoIP services is identified.

Chapter 2 is literature review. We look on different approach to speech quality monitoring and ASR, and the commonality between these two services. The trend from classical signal processing method to Deep Neural network methods is discussed, the DNN performance and applicability depends on training data.

Chapter 3 thoroughly analyses the dataset used to train non-intrusive speech quality prediction models and ASR. Environmental noise and network distortions are not mutually

exclusive. However, there is lack of sufficient studies on mutual effect of environmental noise and network distortion on VoIP service.

Chapter 4 proposes new noise-network distorted speech dataset to facilitate analysis of the mutual effect of environmental noise and network distortions on VoIP application. This chapter discusses the important environmental noise conditions and network condition needs to be captured in the new dataset to reflect on the real environment as much as possible. Next, the process to generation the required speech dataset is discussed. This includes mixing clean speech signal with environment noise, then transmitting the noisy speech signal through emulated network. Emulated network allows controlled change of network condition. On receiver part the transmitted noisy speech is received and decoded to get noise-network distorted speech dataset at different network and noise conditions.

In Chapter 5, analysis work of the effect of environmental noise and network distortion on automatic speech quality monitoring is done. MiniatureVQNet, a single-ended speech quality evaluation method for VoIP audio applications based on a lightweight deep neural network (DNN) model is proposed. This chapter includes the rationale for the proposed model design. The proposed model can predict the audio quality independent of the source of degradation, whether noise or network, or noise and network. The proposed and is light enough to run in embedded systems.

Chapter 6 The MiniatureVQNet model's performance on different VoIP speech degradation factors, including network distortion and acoustic/environmental noise is presented. Finally, to enable the model to run on wide variety of computing systems, different post-training optimization methods were considered to improve the performance in low-resource computing environments such as embedded systems.

Chapter 7 analysis work of the effect of environmental noise and network distortion on automatic speech recognition is done, and proposed mitigation measures are proposed. This chapter presents the experiment setup for performance evaluation of the ASR on noise-network distorted speech dataset. The evaluation is done on two ASR speech models: the ASR model trained on clean speech dataset (CSM) and the ASR model trained on noise-network speech dataset (NNSM). The NNSM was trained by fine-tuning a pre-trained CSM on noise-network distorted speech. The evaluation examines the effects of noise type and network conditions such as delay, jitter and packet loss on the performance of the ASR systems. Then, the performance of the two models is compared on clean speech dataset, and noise-network distorted speech dataset.

Finally, chapter 8 discusses the general outcome of this study, contributions and suggests future research possible research works which were not covered by this study.

CHAPTER 2 LITERATURE REVIEW

Automating speech quality monitoring in VoIP applications can help to ensure that users have a positive experience, free from interruptions and distortions. Automatic speech quality monitoring can detect and manage factors that affect speech quality in real-time, providing a more reliable and consistent service. ASR technology can be integrated into VoIP systems to automate the transcription of spoken language into text. This can be useful for various purposes, such as transcribing voice messages, generating subtitles, or enabling real-time closed captioning during live events. This chapter discusses different methods of automating speech quality monitoring and ASR, and what measures are taken to mitigate the effect of noise and network distortions in VoIP applications.

2.1 Speech Quality Evaluation

Since the invention of the telephone, voice-based telecommunication has been widely adopted. Audio communication involves audio signal capturing, processing, and transmission. The receiving end processes the received signal to output the reconstructed captured/original audio signal. This process introduces audio degradation. The introduced degradation affects the quality of service offered to users. Therefore, there is high interest in monitoring audio quality, mainly when the audio is transmitted over an Internet protocol (IP) network, the most commonly used global network. The IP network is a best-effort delivery network whereby the network does not guarantee that data are delivered or meet any quality of service, as the network performance depends entirely on the traffic load.

Audio quality degradation in IP audio systems is introduced by environmental noise, poor network quality, and audio encoding–decoding algorithms. The quality of audio signals in IP networks depends much on the network's performance at a given time instant. In VoIP applications such as phone calls and conferences involving two-way communications between humans, when the audio signal deteriorates, the involved parties can easily notice the situation and take mitigation measures. However, in some IP audio applications, such as IP based broadcasting systems, Internet radio, and music streaming apps, the communication is always one way, from the transmitter to the receivers. There is no way for the transmitter to know the quality at the receiver end. Therefore, there is a need for an automatic quality evaluation of the transmitted audio. Speech quality monitoring in VoIP systems enables autonomous system adaptation, e.g., changing encoding parameters or re-announcing. Furthermore, the IP network is shared with other applications. Hence, monitoring can detect when the network is overloaded and the monitoring results can be

used as an input to an Adaptive Bitrate Switching Algorithm for Speech Applications (Alahmadi et al., 2021), so that coding parameters can be changed based on the network condition.

Currently, voice/speech and music are the major categories of audio transmitted over IP networks. Music and speech share common audio signal characteristics but have different objectives for listeners. High-quality speech and music audio signals should be of the same fidelity as the source after processing and transmission. Moreover, for speech communication, how easy it is to understand words utterances in the received speech audio signal after processing is very important. This measure is called intelligibility.

Speech quality measures assess a speaker's utterance voice quality, including attributes such as natural, raspy, hoarse, scratchy, and so on. Speech quality is highly subjective and difficult to evaluate reliably because individual listeners have different standards of what constitutes good or poor quality, resulting in large variations in rating scores among listeners listening to the same audio. Intelligibility measures assess "what words the speaker said" compared to the "words the listener heard"; that is, the meaning or the content of the spoken words understood by the listener.

Unlike speech quality, speech intelligibility is not subjective. It can be easily measured by presenting speech material (sentences, words, etc.) to a group of listeners and asking them to identify the words spoken. Intelligibility is quantified by counting the number of words or phonemes identified correctly. The relationship between speech intelligibility and speech quality is not fully understood. This is partly because we have not yet identified the acoustic correlations between quality and intelligibility (Voiers, 1980). Speech can be highly intelligible yet be of poor quality.

Speech audio quality assessment can be performed using subjective listening tests or objective measures. In VoIP quality of service the monitored quality is always speech quality not intelligibility, which is also the center of the monitoring part of this work.

2.1.1 Subjective Speech Quality Evaluation Methods

Mean opinion scores (MOSs) are the most widely used direct subjective speech quality evaluation method. The MOS is a categorical judgment method in which listeners rate the quality of the speech test signal using a five-point numerical scale, with five indicating "excellent" quality and one indicating "unsatisfactory" or "bad" quality. The measured quality of the speech test signal is obtained by averaging categorical scores from all listeners, and the average score is commonly referred to as the MOS. This method is recommended by the IEEE Subcommittee on Subjective Methods (IEEE, 1969) as well as by ITU (INTERNATIONAL TELECOMMUNICATION UNION, 1996). In addition, a

subjective test methodology for evaluating speech in communication systems that include a noise suppression algorithm (INTERNATIONAL TELECOMMUNICATION UNION, 2003) , which measures the perceived speech quality, is one of the most adopted methods.

MOS is based on Five-point assessment scale, where the highest numerical score is 5 and the lowest numerical score is 1. The highest score of 5 corresponds to very good speech quality, the score of 4 corresponds to good quality, the score of 3 corresponds to Fair, the score of 2 corresponds to poor speech quality, and the score of 1 corresponds to bad speech quality.

Table 1 MOS Five-point assessment scale

Numerical score	Quality
5	Very good
4	Good
3	Fair
2	Poor
1	Bad

VoIP service providers use subjective speech quality measurement to improve their service. Most service providers ask for the user's feedback after the call has ended. This is often appears as a pop-up message after the call has ended, as shown in Figure 3 VoIP subjective call quality measure by VOIP providers from left: Microsoft teams, Whatsapp, Skype.

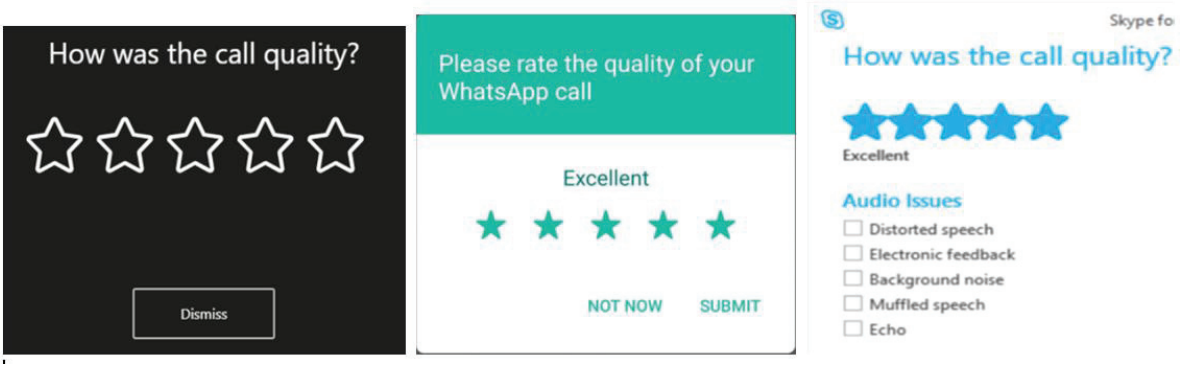

Figure 3 VoIP subjective call quality measure by VOIP providers from left: Microsoft teams, Whatsapp, Skype.

However, subjective speech quality measurement is expensive as it requires the recruitment of human subjects. Furthermore, subjective quality measurement is not practically applicable in automatic speech quality monitoring in real-time VoIP application because it is expensive, not viable to recruit listeners for each established communication, and for security reasons (communication privacy). Therefore, objective quality measures are developed to model subjective listening tests and mathematically estimate the audio quality perceived by human beings. Objective measures enable automation of speech quality estimation task. There are two main categories of objective quality measurement methods: end-to-end (intrusive) and single-ended (non-intrusive) methods.

2.1.2 End-to-End Objective Speech Quality Evaluation Methods

End-to-end objective speech quality evaluation involves the usage of mathematical signal processing techniques to compare the VoIP transmitted/processed speech signals to the original speech signal. Speech objective measures calculate the speech quality by measuring the numerical "distance" between the VoIP transmitted/processed speech signal and the original speech signal. Currently, the widely used end-to-end automatic speech evaluation methods are:-

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). PESQ is the earliest ITU-T recommended speech quality objective measurement method. PESQ was succeeded by Perceptual objective listening quality assessment (POLQA) (Thilo et al., 2000) (INTERNATIONAL TELECOMMUNICATION UNION, 2011) which have better accuracy than PESQ. POLQA has undergone several improvements with the recent approved Perceptual objective listening quality prediction (POLQP) (ITU, 2018) succeeding the POLQA.

These recommended methods had demonstrated acceptable accuracy when the speech is affected by packet loss and packet loss concealment with code-excited linear prediction (CELP) codecs, acoustic noise in sending environment effect, transmission channel errors, effect of varying delay in listening only tests and transcoding, among other factors. However, they cannot be used in in-service non-intrusive measurement devices, two-way communications performance, acoustic noise in receiving environment, echo, and the effect of delay in conversational tests. These methods have been widely applied for designing and evaluating the audio encoding and decoding algorithms.

Other End-to-End objective speech quality evaluation methods, includes the PEMO-Q (Huber & Kollmeier, 2006), and a New Method for Objective Audio Quality Assessment

Using a Model of Auditory Perception. Furthermore, the perceptual evaluation of audio quality (PEAQ) is used to evaluate speech and other audio signals such as music.

End-to-end methods for objective audio quality evaluation require a clean (origin) reference signal and received (VoIP transmitted)/processed speech signal to evaluate the speech quality. End-to-end methods are also called intrusive speech quality evaluation methods because they require reference signal. End-to-end methods dependency on reference signal, limits the method usability in general applications, as, in most cases in VoIP applications, the original speech (reference) signal is not readily available. The performance of IP networks changes based on the network load, affecting the quality of speech transmitted through the network. In the speech quality monitoring of VoIP applications, the interest is in continuously monitoring the quality of the transmitted speech signals. We only have access to the received signal at the monitoring point, whereas the original speech signal is unavailable. Single-ended objective speech quality measures estimate the speech quality using only received/processed signals without needing the original signal.

2.1.3 Single-Ended Objective Measures of Speech Quality

A single-ended objective measure of speech quality is a technique for assessing the quality of speech signals without requiring a reference signal. This type of measure is also referred to as non-intrusive method, as it does not require additional information beyond the degraded signal to make audio quality judgments. Single-ended objective measures of speech quality are useful for evaluating the quality of speech signals in situations where a reference signal or subjective ratings from human listeners are not available or practical. These measures can be used to provide real-time feedback on the quality of speech signals in VoIP communication systems. Single-ended objective measure of speech quality is a practically viable method for continuously monitoring the quality of speech delivered to a VoIP endpoint or a particular point in the network. Based on continuous quality assessment results, network traffic can be re-routed through a less congested route, or codec parameters can be changed, improving the service quality. Different single-ended methods have been proposed based on signal processing techniques or machine learning approaches.

Single-ended (non-intrusive) quality measures are suitable for continuously monitoring the received speech quality in IP networks, as there is no need for an original/reference signal. The IP network is the widely used global network implemented in Internet networks, local area networks (LANs), and enterprise networks. The wide availability of IP networks has driven the integration of voice/audio communication into the IP network. Voice communication services such as voice calls, conference systems, music streaming, and IP public addressing speakers are offered through IP networks. Therefore, it is desirable that

single-ended speech quality objective measures to correlate well with subjective listening test MOS results to be practically applicable in these applications. However, signal-processing-based single-ended speech quality evaluation results do not correlate well with subjective MOS test results.

Clearly, it is desirable that objective measures correlate well with subjective listening test MOS results. However, signal processing based single-ended speech quality evaluation results does not correlate well with subjective MOS test results. Signal-processing-based methods such as single-ended methods for objective speech quality assessments in narrow-band telephony applications (P.563) (INTERNATIONAL TELECOMMUNICATION UNION, 2004) have been outperformed by the recently proposed deep learning methods [(Sharma et al., 2016), (Gamper et al., 2019), (Cauchi et al., 2019), (Catellier & Voran, 2020)]. However, deep learning models depend much on the training dataset used. Previous proposed deep-learning-based single-ended methods [(Gamper et al., 2019), (Cauchi et al., 2019), (Catellier & Voran, 2020)] were limited to acoustic characteristics. In (Sharma et al., 2016), the effect of telecommunication networks were considered using speech transmitted through cellular networks and telephone networks. However, cellular and telephone networks exhibit different properties from IP networks.

Furthermore, the diversity of IP audio transmitters and receivers is high, from high-performance personal computers and mobile phones to low-memory and low-computing-capacity embedded systems. Therefore, deep learning methods have an excessive resource demand, impeding its deployment on low-end devices. Ideally, the objective speech quality measure in VoIP applications should be able to estimate the quality independent of the type of speech distortions introduced by the VoIP system, whether network distortions, speech encoding–decoding, or environmental noise. Moreover, the objective quality measure should support as diverse devices as possible.

2.2 Automatic Speech Recognition

Automatic speech recognition (ASR) systems provide services such as voice search, voice command, and automatic call transcription. Thus, ASR systems have been widely implemented in health care systems, virtual assistants on mobile devices, and cognitive bots. Because speech is the most preferred and natural mode of communication between humans, the industries utilizing ASR applications will continue to expand. To improve the user experience and quality of service, the ASR is built into VoIP applications. Few examples include artificial intelligence (AI) powered meeting transcription and transcription in call centers. The VoIP-transmitted speech presents a new challenge to the ASR systems as it originates from diverse sources and is captured in varying levels of

environmental noise. This encompasses hand-free devices on cars and VoIP calls in noisy environments such as train stations and airports among other examples. Furthermore, VoIP speech signals are distorted not only by the environmental noise but also the transmission network. These characteristics of VoIP speech signals cause hindrance to the designing of robust and high accurate ASR systems. Deep learning has outperformed other ASR techniques (Li & Sim, 2014). To build highly accurate and robust deep-learning-based ASR, various techniques has been studied, including feature extraction, language models, deep learning architectures, and rich characteristics datasets(Malik et al., 2021). Deep learning is reliant on the availability of massive amounts of data. Thus, speech datasets for the development and evaluation of ASR systems have evolved over time.

2.3 Speech Quality Evaluation and ASR in VoIP

Speech quality evaluation is an important aspect of Voice over Internet Protocol (VoIP) systems, which transmit voice data over the internet. ASR is also a critical component of VoIP systems, as it allows for automatic transcription of spoken words into text, which can be useful for call transcription, analysis, and search. These two components of VoIP system share one main common aspect; they are all heavily dependent on machine learning methods, particularly deep learning methods. Training deep learning models is reliant on the availability of massive amounts of data. In the next chapter, we will look on different datasets used for training ASR models and non-intrusive speech quality evaluation models.

CHAPTER 3 NON-INTRUSIVE SPEECH QUALITY MONITORING AND ASR SYSTEMS TRAINING DATASETS

Deep-learning-based models require a large amount of speech data to achieve high accuracy. The accuracy of a deep learning-based model is highly dependent on the features of the dataset used for training. Overall, it is important to carefully consider the features of the dataset used for training an ASR model to ensure that the resulting model is accurate and robust in a variety of real-world scenarios. Acoustic noise and network distortions are inherent part of VoIP applications. Therefore, acoustic noise and network distortions features needs to be well presented in dataset for training ASR and Quality monitoring systems. If acoustic noise and network distortion distortions features are not well covered in the training dataset, there is a risky of the deep-learning models to have poor performance in real-world application.

3.1 Non-intrusive Speech Quality Monitoring Training Datasets

Non-intrusive speech quality monitoring training dataset can be categorized into two groups. The first group is models that evaluate the quality of speech in VoIP applications based on a set of transmission parameters without processing the actual received speech signal. In parametric models there is no need for the reference signal, but it is not a truly non-intrusive, as the model input is parameters obtained from the transmission network. The second group is a group of models that depends on the received speech as the only input for speech quality prediction.

3.1.1 Parametric Models for Speech Quality Evaluation

Parametric models evaluate the speech quality from a set of parameters without processing the actual audio signal samples. Parametric models are commonly used in VoIP applications and are derived from models such as the E-model(ITU-T, 2015). The E-model takes into account several factors that affect the quality of a voice call, including network impairments, such as delay, jitter, and packet loss, as well as voice codec characteristics and endpoint equipment characteristics. The model uses these factors to predict the R-factor, which represents the expected mean opinion score (MOS) of a voice call.

The E-Model is the most extensively applied parametric objective assessment method; it was originally designed for conventional network planning. The E-model has been widely

adopted as a standard for predicting the speech quality of IP-based communication systems, and it is used by many network operators, service providers, and equipment manufacturers to design and optimize their systems for optimal speech quality.

Recently, machine learning speech quality prediction models have better performance than mathematical models such as the E-model. Machine learning based parametric models evaluate the quality of experience (speech quality experienced by VoIP users) in VoIP networks based on network quality of service (QoS) mapping and machine learning algorithms. [(Sun & Ifeachor, 2002a), (Sun & Ifeachor, 2006), (Rodriguez et al., 2019a), (Z. G. Hu et al., 2020)].

3.1.2 Parametric Models Dataset and Weakness

Parametric model studies are primarily concerned with the effect of network conditions on the speech quality experienced by users. The network packet loss effect is the most studied.

The method proposed in (Wuttidittachotti & Daengsi, 2017) evaluated the influence of packet loss on the Skype quality and proposed a simplified E-Model. In (Jelassi & Rubino, 2018), the authors addressed the effect of burst packet losses on VoIP.

In (Uhl, 2018), the effect of different codec and network conditions on the speech quality was examined. In (Rodriguez et al., 2019b), work was carried out to improve the E-model for wireless communication systems based on wireless parameter values and consider current technologies, such as MIMO. Moreover, an approach that considered the interactivity of voice communications was developed by Sun and Ifeachor (Sun & Ifeachor, 2006), where the E-Model and PESQ were combined to evaluate the voice quality, and a nonlinear regression model of voice quality was proposed.

Parametric models can evaluate distortions due to network conditions and devices used. However, as they do not analyze the actual audio samples, if the samples are distorted due to environmental noise, parametric models cannot be applicable for evaluating speech quality. Parametric models are not truly non-intrusive as the prediction depends on pre-calculated device characteristics and intrusion of network parameters.

3.1.3 Non-intrusive Models Speech Dataset

A wide variety of datasets are used to train DNN-based non-intrusive speech quality prediction models. For speech enhancement algorithms, artificially added environmental noise is commonly used. In addition, noise and speech transmitted through cellular networks, or a combination of the two, are used to assess the distortions of telecommunication networks.

Table 2 A summary of DNN method for non-intrusive speech quality evaluation and dataset features used for training the models

Proposed method	Dataset	Acoustic noise	PTSN	VoIP	Network conditions
MetricNet (Yu et al., 2021)	AISHELL-2 corpus Room simulation	reverberant and noisy	×	×	×
Aecmos: a Speech Quality Assessment Metric for Echo Impairment (Purin et al., 2022)	Self-generated dataset	Noisy	×	×	×
Deep autoencoder (Soni & Patil, 2016a)	NOIZEUS database	noisy	×	×	×
Modulation Energies and LSTM-Network (Cauchi et al., 2019)	Noise and reverberation simulation	reverberant and noisy	×	×	×
support vector machine (SVM) classifier for speech evaluation (Islam et al., 2017)	NOIZEUS database	noisy	×	×	×
Activity detection and entropy based, ASR network (Ooster & Meyer, 2019)	WSJ1 speech corpus and Aurora4 maskers as additive noise	noise	×	×	×
Quality-Net (Fu et al., 2018)	TIMIT corpus	noisy	×	×	×
Convolution Neural network model (Gamper et al., 2019)	Noise and reverberation simulation	reverberant and noisy	×	×	×
NSQM (Jassim & Zilany, 2019)	NOIZEUS database	noise	×	×	×
Wavenets (Catellier & Voran, 2020)	Open source datasets	Noise	×	×	×
NISA (Sharma et	TIMIT database	noisy	Yes	×	×

al., 2016)	NATO database CTIMIT database NTIMIT database C-Qual database				
DNN No-Reference PSTN Speech Quality Prediction (Mittag et al., 2020)	automated call using Librivox database	×	Yes	Yes	×
NIML (Alkhawaldeh et al., 2019)	Network condition simulation	×	×	Yes	Yes
artificial neural network (ANN) model (Sun & Ifeachor, 2002b)	TIMIT packet loss	×	×	Yes	Yes
NISQA (Mittag & Möller, 2019)	29 different databases	Noise	Yes	Yes	×
Tree-CNN (Vieira et al., 2020)	Noise and Packet loss simulation	noise	×	Yes	Yes
GMM-Based (Falk & Chan, 2006b)		reverberant and noisy	Yes	Yes	×

These databases contain a large variety of speech distortions, such as different codecs, noises, live recordings, and transmission errors. Dataset used for training machine learning based speech quality prediction models can be grouped into noise only, network only and dataset that combines noise and network effect. In Table 2, we have analyzed commonly used datasets based on four criteria:-

3.1.3.1 Acoustic Noise

This includes all speech data that contains noise. The noise can be either artificially added to clean recorded speech or the speech was recorded in a noisy environment. The recording settings can be single microphone capture or multi-channel microphone. The, noise type includes reverberation and surrounding noise. This feature is very common in many datasets, the fact that it can be used to train speech enhancement algorithms and also speech quality prediction algorithms. Type of noise and SNR information is label in the dataset.

3.1.3.2 Public Switched Telephone Network

In this case we looked on dataset that was transmitted through PSTN network. In PTSN network microphone captured speech signal is encoded, transmitted and then decoded. This process induces encoder and network transmission errors. We also looked to see if the network transmission parameters were included in the dataset or not. In real phone calls, environmental noise cannot be completed isolated. Therefore, we looked to see whether the transmitted speech signals were clean signal or noise speech signal. There are dataset that contains network transmission errors only and other datasets with a combination of network transmission errors and acoustic noisy.

We found few cases where speech signal transmitted through PTSN network was used to train speech quality prediction algorithms. The NISQA (Mittag & Möller, 2019) and GMM-Based (Falk and Chan 2006a) dataset, this dataset contains speech signals with different acoustic noise, speech transmitted through PTSN network and also speech signals transmitted through VoIP network. Looking at VoIP application, this dataset covers all the possible scenarios that of environmental noise and network related induced distortions. However, this dataset does not contain information on parameters of the VoIP network of the audio files. Hence it is difficult to analyze the individual effects on VoIP network parameters on performance of the prediction model.

3.1.3.3 VoIP Distortions

The third, feature we looked on the dataset was weather the dataset speech signal contains VoIP distortions. VoIP and PTSN network was grouped differently, because of the inherently different of PTSN network and IP network. In VoIP network there are two cases, where the network through which the speech signal was transmitted is real internet or simulated-network environment. In the real IP network is difficult to get the precise condition of the network because the speech signal is routed through different administrative networks, however for simulated environment network condition parameters can be controlled.

NIML (Alkhawaldeh et al., 2019) and artificial neural network (ANN) model (Sun & Ifeachor, 2002b) models were trained using VoIP transmitted speech on simulated network environment. Therefore these dataset contains transmitted network parameters, however these dataset does not contain environmental noise feature. Hence, they cannot be used as a real representative of VoIP application conditions. GMM-Based (Falk & Chan, 2006b) dataset contains both environmental noise and IP network transmission distortion. However the speech dataset does not contain information on network for each speech utterance.

3.1.3.4 VoIP Network Condition

We looked if the dataset includes IP network condition parameters through which the audio speech was transmitted. Many dataset includes network condition parameters, but they do not include noise distortion, it is clean transmitted speech and controlled network parameters. Tree-CNN (Vieira et al., 2020) dataset included acoustic noise, VoIP transmitted speech with network conditions. However, the only network feature included is packet loss. This dataset does not include other important IP network parameters such as jitter, delay, and bandwidth.

There are different datasets used in training speech quality prediction models, however, the majority of the datasets includes on speech degraded by environmental noise and reverberation. In case of VoIP speech degradation measures, the most of the datasets concentrate of IP network packet loss, and no other parameters like delay, jitter, and bandwidth. However, in VoIP application speech quality degradation is mutual effect of acoustic noise and network distortion.

3.2 Mutual Effect of Acoustic Noise and Network Distortion Datasets in Speech Quality Prediction

Network distortions and environmental noise distortions affect speech quality. Speech signals carry the effect of both noise and network distortions. Models trained on noise-network-distorted speech signals can learn to predict the speech quality regardless of whether the speech is affected by acoustic noise, network distortion, or both.

In (Z. G. Hu et al., 2020) , a model trained and evaluated on real PSTN call recordings from 80 providers in more than 50 countries was proposed. After going through a gateway, the signals were routed across one of five PSTN carrier networks that Skype and Microsoft Teams use. Background noise was artificially added in the call to model environmental noise. In (Mittag et al., 2020), an open-source PSTN speech quality test model based on a dataset with over 1000 crowd-sourced real phone calls was presented. The influence of file cropping on the perceived speech quality and the influence of the number of ratings and training size on the model accuracy were analyzed. The proposed models in [(Z. G. Hu et al., 2020), (Mittag et al., 2020)] were trained on a public network and crowd-sourced data, and the distortion parameters such as packet loss and environmental noise were not controlled. Therefore, the speech quality distortions could not be independently analyzed, and the model accuracy in different conditions could not be verified. In this work, in

addition to the public available dataset, we generate new dataset in a controlled environmental where all factors are known and can be analyzed.

3.3 ASR Training Datasets

3.3.1 Studio-Recorded Read Speech Dataset

Traditional deep-learning-based ASR systems were trained on read-speech studio-recorded read speech signal datasets. Read speech is the speech utterances from a written script, where a reader reads the script as it written. In the studio environment noise is isolated and can't interfere with the intended speech signal. Studio-recorded read speech datasets is clean speech without any environmental noise.

The studio-recorded read speech dataset was used in early ASR works. The earlier such datasets are the ATR Japanese speech database (Kurematsu et al., 1990), TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993), WSJ corpus (Charniak et al., 2000), and Mandarin Chinese broadcasting news (Wang et al., 2005). Studio-recorded read speech lacks the naturalness of human conversation, for example, readers do not think what to say. In natural communications people tend to spontaneously utter speech based on topic, mood, etc.

3.3.2 Studio-Recorded Spontaneous Speech Dataset

Studio-recorded read speech dataset was successful in arousing the interest in the study of ASR. However, a speech read from a pre-prepared script lacks the naturality of everyday human speech. Furui et. al. introduced a Corpus for Spontaneous Japanese (CSJ) (Furui et al., 2000). CSJ is a database containing a large collection of Japanese spoken language data and information for use in linguistic research.

The CSJ corpus has been used for a wide variety of research purposes such as spoken language processing, natural language processing, phonetics, psychology, sociology, Japanese education, and dictionary compilation. The spontaneous monologue bears close resemblance to the natural human conversation. With the prevalence of deep learning, much larger datasets were introduced to tap into the potential of deep learning in ASR.

3.3.3 Large Dataset Size

Deep learning based ASR models trained on large datasets tend to yield high performance. Large datasets such as LibriSpeech (Panayotov et al., 2015) LibriSpeech is a corpus of

approximately 1000 hours of 16kHz read English speech. TED-LIUM Corpus (Rousseau et al., 2012), (Köhn et al., 2016) and Common Voice(Ardila et al., 2020) contains a thousand hours of speech. These large datasets have improved the performance of ASR systems. However, the performance still degrades in real application environments, where speech signals are usually captured with environmental noise.

3.3.4 Speech Recorded in Natural Environment

The ASR systems trained on large datasets of studio-recorded speech exhibit low performance on noisy or degraded speech. This has led to the introduction of speech datasets recorded in natural environments such as the domestic setting, for example the DIRHA-English corpus (Ravanelli et al., 2015), CHiME-2 (Barker et al., 2013), CHiME-3 (Barker et al., 2015), and CHiME-5 (Barker et al., 2018).

Natural environment speech dataset encompasses recorded speech spoken live in noisy environments and simulated speech datasets that were generated by artificially mixing the clean speech data with noisy backgrounds. The introduction of speech datasets recorded in natural, noisy environments has improved the performance of ASR systems on noisy speech. However, the speech may also get distorted as a result of degradation that occurs when it is transmitted through an IP network.

3.3.5 Speech Distorted in Transmission Network

Although Environmental noise degrades the speech quality, distortion and degradation are also introduced in the transmission of speech through computer networks. Low bandwidth, echo, encoding–decoding distortion, differences in handsets, and network poor quality all present new challenges toward building robust and highly accurate ASR systems.

For the CTIMIT (Brown & George, 1995) dataset generated by transmitting clean voice speech through a cellular network, network distortions caused a 58% drop in the ASR performance. Training on network-distorted speech increased the recognition accuracy by 82%. VoIP applications use packet-switched networks, which have different characteristics from that of circuit- switched networks.

In the VoIP applications, the speech quality is degraded by delay, jitter, packet loss, packet burst loss, network bandwidth, encoding and decoding algorithms (da Silva et al., 2008) . The effect of combined noise, network and encoding parameters on deep-learning-based ASR models has not been extensively studied.

3.3.6 Acoustic Noise and Network Distorted Speech Dataset

Table 3 Speech datasets used for training ASR models and their characteristics

ASR model	Dataset	clean	noisy	PTSN	VoIP	Network conditions
ATR Japanese speech database as a tool of speech recognition and synthesis (Kurematsu et al., 1990)	ATR Japanese speech database (Kurematsu et al., 1990)	Yes	×	×	×	×
RNNDROP (Moon et al., 2016)	TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993)	Yes	×	×	×	×
	WSJ corpus (Charniak et al., 2000)	Yes	×	×	×	×
	Mandarin Chinese broadcasting news (Wang et al., 2005)	Yes	×	×	×	×
	Corpus for Spontaneous Japanese (CSJ) (Furui et al., 2000)	Yes	×	×	×	×
The Kaldi speech recognition toolkit (Panayotov et al., 2015) (Povey et al., 2011)	LibriSpeech (Panayotov et al., 2015)	Yes	×	×	×	×
	TED-LIUM Corpus (Rousseau et al., 2012) (Köhn et al., 2016)					
	Common Voice(Ardila et al., 2020)					
	DIRHA-English corpus (Ravanelli et al., 2015)					
	CHiME-2 domestic environment (Barker et al., 2013)	Yes	Yes	×	×	×
The NTT CHiME-3 system(Yoshioka et al., 2015)	CHiME-3 (Barker et al., 2015)	Yes	Yes	×	×	×
	CHiME-5 (Barker et al., 2018)	Yes	Yes	×	×	×
	CTIMIT (Brown & George, 1995)	Yes	×	Yes	×	×

Large variety of speech dataset has been used to train ASR system models. The dataset contains a combination of features such as such as different codecs, noises, live recordings, and transmission errors. We have examined the dataset used for training ASR system based on four main features noise only, network only and dataset that combine noise and network effect, which are the features of concern in this work. Table 3, shows the dataset and speech features captured in the dataset.

3.3.6.1 Clean Speech Dataset

This is a common type of speech dataset used for training ASRS systems. The speech signal is recorded in studio environment. Therefore, the recorded speech signal does not include environment noise. This early ASR systems models (Kurematsu et al., 1990) (Garofolo et al., 1993) were trained using this type of dataset.

3.3.6.2 Noisy Speech Dataset

Background noise is a natural inherent part of the environment where VoIP systems are used. VoIP systems are not used only in studio like environment where background noise is controlled. Therefore, it is important for ASR systems to be trained using noisy speech datasets.

To create noise speech dataset, noise can be either artificially added to clean recorded speech or the speech can be recorded in a noisy environment setting. The commonly used noise environment settings are train stations, restaurant, airport, etc. The recording settings can be single microphone capture or multi-channel microphone CHiME-5 (Barker et al., 2018). The, noise type includes reverberation and surrounding noise. Noisy dataset is common used in speech quality recognition systems, but in training ASR systems their use has started recently (Barker et al., 2013), CHiME-3 (Barker et al., 2015), CHiME-5 (Barker et al., 2018) .

3.3.6.3 Public Switched Telephone Network

In this case we looked on dataset that was transmitted through PSTN network. In PTSN network microphone captured speech signal is encoded, transmitted and then decoded. This process induces encoder and network transmission errors. We also looked to see if the network transmission parameters were included in the dataset or not.

In real-life phone calls, environmental noise cannot be completed isolated.

Therefore, we looked to see whether the transmitted speech signals were clean signal or noise speech signal. There are dataset that contains network transmission errors only and other datasets with a combination of network transmission errors and acoustic noisy.

Cellular TIMIT (CTIMIT) (Brown & George, 1995), has been generated by transmitting the TIMIT speech database over the cellular network, with the aim of aiding in the design and development of speech processing and speech recognition products for the cellular market. The ASR system trained on clean TIMIT dataset experiences a 58% decrease in phonetic recognition accuracy when tested with the CTIMIT test corpus. By contrast, the ASR trained on CTIMIT the accuracy increased 82% compared to that of the TIMIT-trained recognizer.

3.3.6.4 VoIP Distortions and VoIP Network Conditions

The fourth, feature we looked on the dataset was weather the dataset speech signal contains VoIP distortions. VoIP and PTSN network was grouped differently, because of the inherently different of PTSN network and IP network. In VoIP network there are two cases, where the network through which the speech signal was transmitted is real internet or simulated-network environment.

In the real IP network is difficult to get the precise condition of the network because the speech signal is routed through different administrative networks, however for simulated environment network condition parameters can be controlled. There are few cases where VoIP and network distorted speech datasets were used to train non-intrusive speech recognition systems

There are different datasets used for training ASR systems, however, the majority of the datasets includes on speech degraded by environmental noise and reverberation. In case of VoIP speech degradation measures, the most of the datasets concentrate of IP network packet loss, and no other parameters like delay, jitter, and bandwidth. However, in VoIP application speech quality degradation is mutual effect of acoustic noise and network distortion.

3.3.7 Mutual Effect of Acoustic Noise and Network Distortion Datasets in Automatic Speech Recognition

In VoIP applications environmental noise and network distortions can neither be completely isolated nor isolate. Network distortions and environmental noise distortions affect speech quality. Speech signals carry the effect of both noise and network distortions. The effect Speech quality prediction models are more robust than those trained on clean.

There are few studies of the mutual effect of noise and network distortion on speech signals input to ASR systems. Brown's (Brown & George, 1995), study compared the performance of ASR system trained on clean speech signal and that of the ASR system trained on clean speech signal transmitted through cellular network. The ASR system trained on network distorted speech dataset has higher prediction performance in cellular application compared to the ASR system trained on clean speech dataset. However there is no such study for the mutual effect of noise and network distortions.

CHAPTER 4 NOISE–NETWORK–DISTORTED SPEECH DATASET

In VoIP applications, speech is distorted by captured environmental noise, encoding–decoding, and transmission network distortions. There is no public available speech dataset contains noise distortions, network distortions and information on noise and network distorting parameters. To examine mutual effect of noise distortions and network distortions on speech quality prediction and automatic speech recognition we had to create new dataset for this task. The newly created dataset for the integrated effect of environmental noise, encoding–decoding, and transmission network distortions used to study the performance of ASR systems and speech quality prediction models is referred as noise–network–distorted speech dataset.

The dataset was generated by artificially corrupting the clean speech dataset with different environment noise at different signal-to-noise ratios and then transmitting the noise-mixed speech signal through the emulated network. As a result, the distorted noise–network dataset contains information on environmental noise conditions such as signal-to-noise ratios and noise types, and network condition parameters such as packet loss, delay, and jitter, as the data were generated in a controlled environment.

4.1 Noisy Speech Properties

Noisy speech signals were obtained from NOIZEUS noisy speech corpus (Y. Hu & Loizou, 2007). This database contains 30 IEEE sentences produced by three male and three female speakers, recorded in a sound-proof booth, and then artificially corrupted by eight different real-world noises. The noises are Babble, Car, Exhibition Hall, Restaurant, Street, Airport, Train Station, and Train. The NOIZEUS noisy speech database includes all the phonemes in the American English language. The sentences were originally sampled at 25 KHz and downsampled to 8 kHz. The clean speech signals were corrupted by adding noise at SNRs of 0 dB, 5 dB, 10 dB, and 15 dB.

4.2 Network Characteristics

VoIP network QoS characteristics include delay, jitter, bit rate, loss rate, and loss distribution. VoIP application parameters that affect speech quality can also include Encoding–decoding parameters, such as bit rate and Forward error correction. The impact of these factors on the perceived QoS in VoIP communications has been studied

extensively (Sun & Ifeachor, 2002a),(Sun & Ifeachor, 2006), (da Silva et al., 2008) , (Z. G. Hu et al., 2020).

There are many network QoS parameters that can affect the transmitted speech signals quality. However, we considered only those with a high impact on QoS of VoIP applications, which are: loss rate, burst packet loss, delay, and jitter. We studied G.722 (ITU-T, 2005) a wideband speech codec.

Combining the VoIP and acoustic characteristics, the new noise-network distorted speech dataset characteristics are summarized in Table 1.

Distortion	Parameter	Values
Network	Packet Loss (%)	0, 10, 15, 20, 25, 30, 35
	Delay (ms)	0, 100, 200, 300, 500
	Jitter (% delay)	0, 10, 20, 30, 40
	Codec	G722
Noise	Noise type	Babble, Car, Exhibition Hall, Restaurant, Street, Airport, Train Station, Train
	SNR (dB)	0, 5, 10, 15

Table 4 Noise-Network distortion parameters

The dataset also included clean speech audio files; these were studio-recorded utterances with no artificially added noise. Clean speech utterances were also distorted when they were sent through the networks. The parameters were selected to closely match those of the real-world environmental noise and the internet QoS. The noise types were babble, car, exhibition hall, restaurant, street, airport, train station, and train. The network parameters were selected to encompass the characteristics of both good and poor internet quality environments.

4.3 Noise-Network Distorted Dataset Generation

To generate the noise-network distorted speech database, the clean speech artificially corrupted by noise was transmitted through an emulated network.

4.3.1 Network Emulation Configuration

We used Netem (Linux Foundation, 2021) network emulation software in collaboration with Tc (Hubert, 2001) a tool used to configure traffic control in the Linux kernel. Netem and Tc provide network emulation functionalities to emulate the properties of wide area networks. The Netem is a kernel component which can be enabled or disabled. In recent Linux distributions, Netem is already enabled and Tc software is pre-installed. A speech generation environment was set up as shown in Figure 4. The router had the following hardware specifications: SoC Broadcom BCM2837 1.2 GHz ARM Cortex-A53 Quad Core Processor (ARMv8 Family), Memory: 1 GB LPDDR2 running Debian operating system (OS). The Debian OS had Netem and Tc enable Linux kernel. FFmpeg version 4.2.4 (The FFmpeg developers, 2020) the open-source command-line tool for converting multimedia formats was used to encode and decode the speech signals. Encoding and decoding application platform was Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz, 32GB RAM, 2TB drive, running Ubuntu OS version 20.04. Speech signals were transmitted using the User Datagram Protocol.

In wide area networks, parameters such as packet loss, jitter, and delay are random variables. Several mathematical models are used to represent this randomness. For simplicity, we used a normal distribution to generate the following noise-network distorted speech delay function:

X is normally distributed with mean μ and standard deviation σ:

$$X \sim N(\mu, \sigma 2)$$

Equation 1 Network packet loss and delay distribution

,where X is the delay distribution, μ is the mean delay, and σ is the jitter, expressed as the percentage of delay as shown in Table 1. In the case of packet loss, the loss was normally distributed with the mean σ = 0.

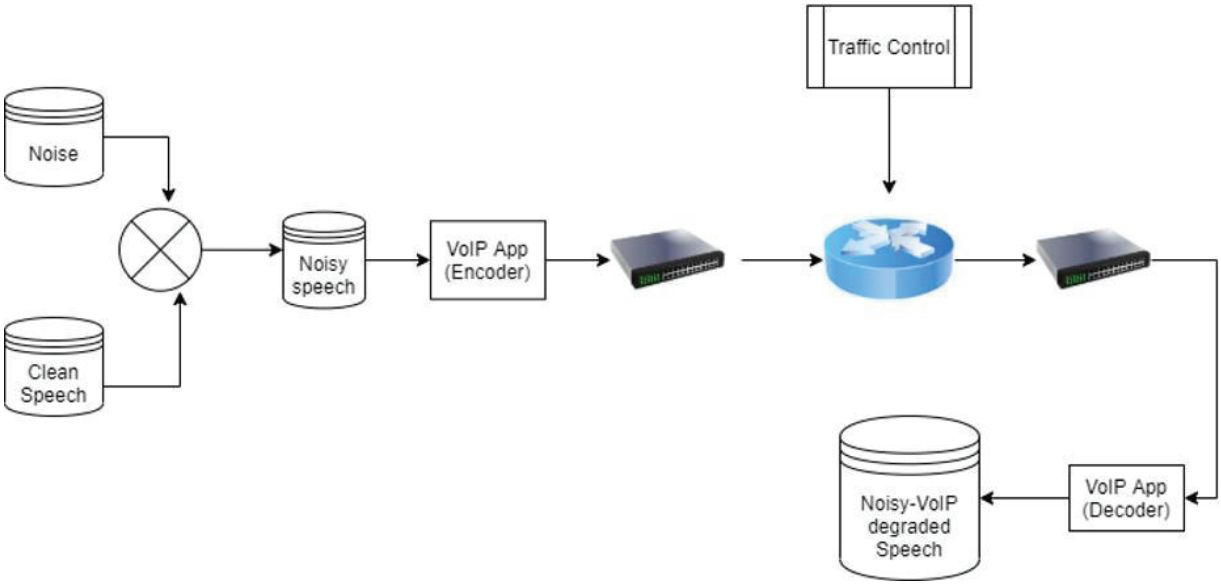

Figure 4 Noise-network speech dataset generation system.

4.3.2 Dataset Generation Process

The noisy speech signals were transmitted by Real-time Transport Protocol (RTP), a network protocol for streaming audio and video over IP networks. A separate TCP channel was used for signaling. The signaling channel was used to indicate the start and end of the transmission, acoustic characteristics of the transmitted speech signal, and network parameters settings in the emulation network used for the transmission. The receiving end decodes the received speech signals, appends the acoustic and network characteristics to create noise-network speech dataset.

4.3.3 Sample Speech Audio Spectrogram for Different Parameters of Noise-Network Dataset

The audio spectrogram for different parameters of noise-network dataset is shown in the Figure 5 below. The clean speech is composed of only speech signal with silence parts. When the street noise at SNR 5 dB is mixed with the clean speech signal, the signal starts to look like white noise. However, the clean speech signal part is easily recognized as it has slightly high power than the noise.

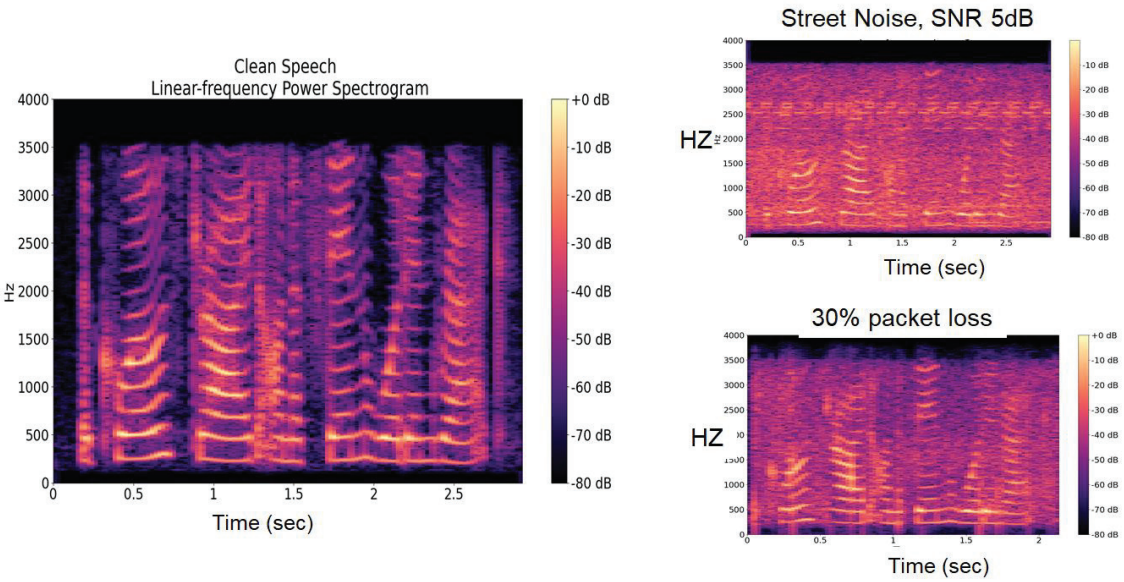

Figure 5 Speech Audio Linear-Frequency Power Spectrogram at Different Noise and Network Condition

In the Figure 5, it is evident that packet loss causes some information loss. The transmitted speech utterance at 30% packet loss network, there is missing information on the delivered speech. The speech length is about 30% shorter than the original transmitted speech.

4.3.4 Generated Noise-Network Dataset

A total of 246,500 sentences with different acoustic and network distortions were generated from 30 utterances. Each sentence length was about 3 to 4 seconds as a result the generated noise-network distortion dataset contained 246.5 hours of speech. The generated noise-network distorted speech database was used to study the performance of the ASR systems for VoIP-based applications.

CHAPTER 5 PROPOSED SPEECH QUALITY PREDICTION METHOD

5.1 Environmental Noise on Speech Quality

In VoIP applications, the microphone that captures speech also captures environmental noise. The environmental noise captured with an intended speech in VoIP affects the quality of the far-end received speech. Several single-ended algorithms for estimating the speech quality of speech signals corrupted by noise have been proposed.

In (Falk & Chan, 2006a) a machine learning method for prediction of speech quality for speech corrupted by multiplicative noise and temporal discontinuities is proposed. The proposed method outperforms the ITU-T P.563, and is capable of measuring the quality of speech enhancement systems.

Deep autoencoder networks have been widely used to capture features in DNN networks. Using deep autoencoder to extract low-dimensional features from a spectrum of the noisy speech signal and finds a mapping between features and subjective scores using an artificial neural network has shown better performance than standard ITU-T P.563 method (Soni & Patil, 2016b). Recently, the MetricNet (Yu et al., 2021), which leverages label distribution learning and joint speech reconstruction learning was proposed. However, all the proposed works does not consider the effect of both noise and network degradations.

In this work, in combination with the noise-network distorted speech dataset network condition parameters, and other public noisy speech datasets, the following noise types were examined: babble (multiple people talking), car, exhibition hall, restaurant, street, airport, train station, and train. The noise signals were artificially added to studio-recorded speech to generate signals at 0, 5, 10, and 15 dBs.

5.2 Effect of VoIP Network QoS on Speech Quality

The network condition in the VoIP system affects the quality of the transmitted speech. IP network conditions change with the network load. Commonly measured IP network conditions for network QoS include delay, jitter, bit rate, loss rate, forward error correction, and loss distribution. There is a substantial body of literature on the impact of network conditions on the perceived speech quality in VoIP communications (Sun & Ifeachor, 2002a), (Sun & Ifeachor, 2006), (da Silva et al., 2008), (Z. G. Hu et al., 2020).

In this study, we examined the effect loss rate, burst packet loss, delay, and jitter, which are the wide studies' network condition parameters in VoIP applications.

5.3 Training Dataset

The training speech dataset was created by combining publicly available datasets and our specific generated noise-network distorted speech dataset in order to cover a wider range of speech quality degradation. The other datasets were:- The noisy speech was taken from a noisy speech database for training speech enhancement algorithms and TTS models (Valentini-Botinhao, 2017), and NISQA Corpus (Mittag et al., 2021).

The speech database for training speech enhancement algorithms and TTS models (Valentini-Botinhao, 2017) was designed to train and test speech enhancement methods, and noise-robust Text-to-Speech models. This speech dataset does not contain any network-induced speech degradations. For each speech utterance in this database, we calculated and labeled the speech quality score using PESQ (INTERNATIONAL TELECOMMUNICATION UNION, 2001) based on the clean reference speech.

Moreover, NISQA Corpus (Mittag et al., 2021), an aggregation of approximately 14,000 speech samples from different datasets, was used. It contains simulated and live phone (mobile phone, Skype, Zoom, Whatsapp) recordings under different noise, networks, and applications such as mobile phone, SKype, Zoom, Whatsapp, and so on. These samples are human-rated and labeled with MOS. NISQA corpus includes VoIP degradations, but the speech files do not contain information on noise and network condition parameters.

To understand the prediction performance on different noise types and network conditions, a simulated noise–network–distorted speech dataset containing information on environmental noise conditions and network condition parameters was used. Environmental noise parameters were different noise types at different signal-to-noise ratios, whereas network parameters were packet loss, delay, and jitter.

The noise–network–distorted speech dataset was divided into training and test datasets. Through the stratified random sampling, 20% of the total sample was set for testing, while the remaining 80% was used for training. The training subset was used for training the single-ended speech quality prediction model in combination with the publicly available datasets, the noisy speech database for training speech enhancement algorithms and TTS models (Valentini-Botinhao, 2017) , and NISQA Corpus (Mittag et al., 2021). The noise–network-distorted speech testing subset was used to analyze the prediction model in different acoustic and network conditions, as this dataset contains acoustic and network condition parameters, whereas the public datasets does not include noisy and network parameters.

5.4 Proposed MiniatureVQNet Model

Early work on machine-learning-based non-intrusive speech quality evaluation was based on Gaussian mixture probability models (GMMs) (Falk & Chan, 2006a) . In (Ooster et al., 2018), a model based on the standard ASR system that combines a feed-forward DNN (which serves as an acoustic model) with a hidden Markov model (HMM) was presented, and the model reached an average correlation of r = 0.87. The DNN model based on a combination of a CNN and RNN (LSTM) (Mittag & Möller, 2019) showed promising results in predicting the quality of super-wideband speech transmission and the impact of packet loss concealment. In (Manocha et al., 2021), the convolutional networks (TCNs) used learned representations while maintaining the temporal structure of the signal for quality evaluation.

5.4.1 Deep-Learning-Based Single-Ended Speech Quality Measures

Several DNN models for non-intrusive speech quality evaluation have been proposed. DNN models were designed for a specific task or general tasks. This section looks into different DNN model architectures and datasets used in training these models.

5.4.2 Model Network Architecture

This study's proposed models have two parts: speech signal feature extraction and quality prediction. The feature extraction part is a modified standard DeepSpeech ASR system, and the prediction part is a stacked layer of two bidirectional GRUs followed by a fully connected layer.

DeepSpeech is an open-source speech-to-text engine that uses a model trained by machine learning techniques based on Baidu's study (Hannun et al., 2014). The DeepSpeech-based technique does not require hand-designed features to model the background noise, reverberation, or phoneme dictionary. Instead, it depends on large amounts of data for training. This was the motivation behind modifying the DeepSpeech model for feature extraction to avoid hand-designed features, as there are no well-known speech features for capturing transmission network distortions.

The major obstacles to implementing deep neural networks (DNNs) on embedded systems are the large model size and many operations needed for inference. However, several model optimization techniques can be applied to DNN models to deploy the models on resource-constrained systems. The commonly used techniques are weight sharing, network pruning,

knowledge distillation, quantization, and designing compact network architectures. This study applied two optimization techniques: designing a compact network architecture and reducing the model precision (model parameters quantization). As a result, we achieved a very light model weight in the proposed network configuration, with only 7,833 trainable parameters and 0 non-trainable parameters.

5.4.3 Network Architecture and Training

The proposed MiniatureVQNet architecture comprises fully connected and bidirec- tional GRU neural network layers with rectified linear unit (ReLu) activation. The network has four fully connected layers with ReLu activation, followed by two bidirectional gated recurrent unit (GRU) layers. The last bidirectional GRU layer feeds to a fully connected layer connected to the output layer. The number of neurons in the first four fully connected layers is given by 32-32-32-8. The number of neurons in the GRU is 8, and the last two fully connected layers is 8-1. The model was trained using stochastic gradient descent with an exponential decay learning rate schedule at an initial learning rate of 0.01, and the decay steps were set to 1000 at the decay rate of 0.9. The training dataset batch size was 64. The Figure 6 below shows the proposed network architecture diagram.

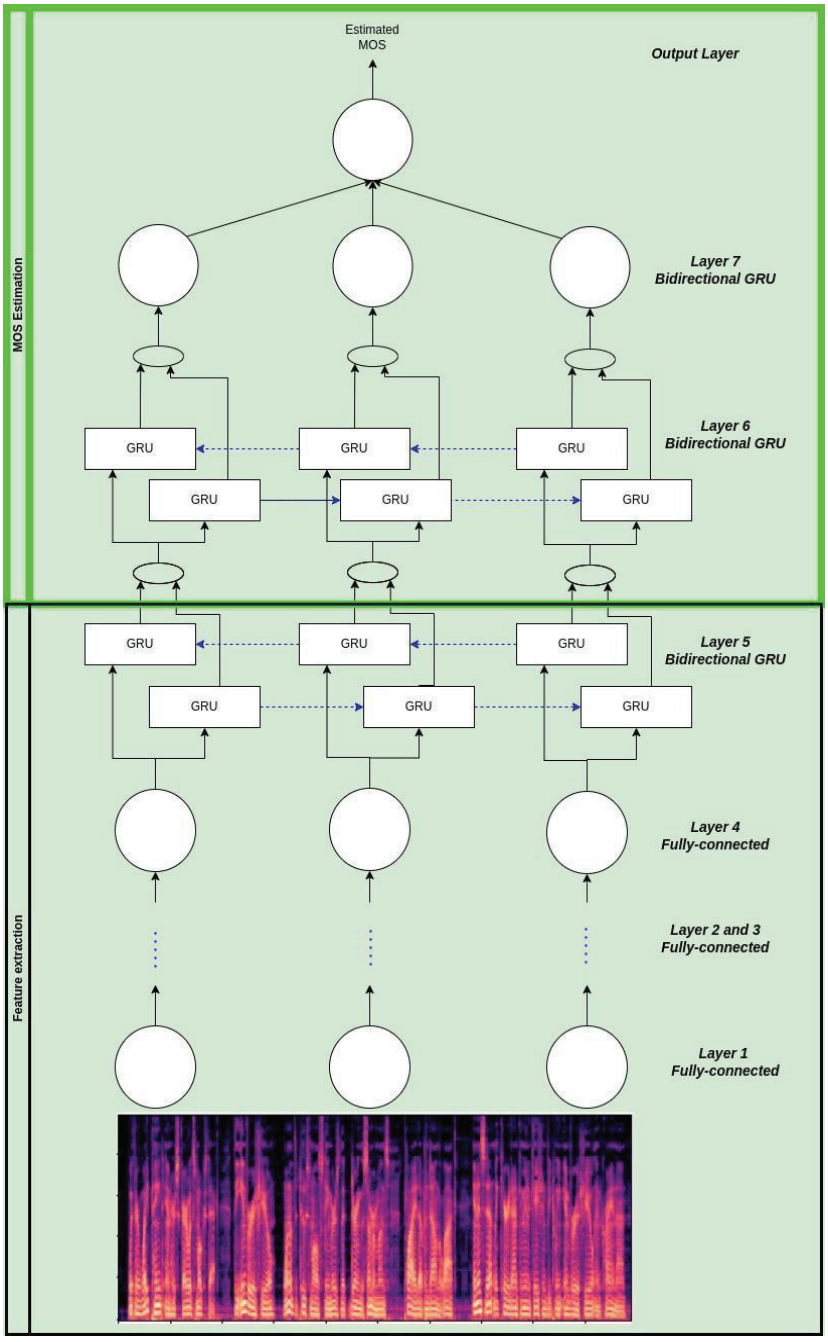

Figure 6 Proposed MiniatureVQNet model network architecture.

5.5 Post-Training Model Optimization

Post-training model optimization is the process of improving a machine learning model after its training phase. This optimization aims to enhance various aspects of the model's performance, such as inference speed, memory usage, accuracy, and robustness to make it more efficient and effective for deployment.

There are several techniques and approaches for post-training model optimization such as quantization, pruning, knowledge distillation, and model quantization-aware training. In this study we considered post-training quantization approach.

Post-training DNN model optimization improves the model performance in a resource constrained environment. Quantization is a post-training model optimization technique that reduces the precision of weights and/or activation functions. As a result, model quantization results in smaller model sizes and faster computation in low-powered devices. The effect of post-training model quantization is reduced model accuracy. However, the benefit of model quantization outweighs the accuracy loss, as Gysel (Gysel et al., 2018) demonstrated that networks could be condensed to use 8-bit dynamic fixed points for network weights and activations with a less than 1% degradation of classification accuracy.

Post-training quantization is a technique to converting the model's weights and/or activations from higher precision data types (e.g., 64-bit floating point) to lower precision data types (e.g., 8-bit integers) without significant loss of performance. Post-training quantization reduces the model size and increases the inference speed, making the model more suitable for deployment on resource-constrained devices or applications that require real-time processing.

In this study, we examined two post-training quantization methods: dynamic-range and float16 quantization. Dynamic-range quantization converts all weights into the nearest 8-bit fixed-point numbers, while the activation and outputs are not changed. Float16 quantization converts the weights and activation to float16. We did not study the Full integer quantization as we did not intent to deploy and test the model on Microcontrollers.

5.5.1 Dynamic range quantization

Dynamic range quantization provides reduced memory usage and faster computation without the need for representative dataset for calibration. Dynamic range quantization statically quantizes only the weights from floating point to integer at conversion time, which provides 8-bits of precision. Dynamic range quantization outputs are stored using floating point.

5.5.2 Float16 quantization

Float16 quantization reduces the size of a floating point model by quantizing the weights to float16, the IEEE standard for 16-bit floating point numbers. When a 32bit floating point model is quantized to Float 16, the model size is reduced up to half. Float 16 quantization supports some delegates (e.g. the GPU delegate) which can operate directly on float16 data, resulting in faster execution than float64 or float34 computations.

5.5.3 Post-training Quantized Models' Size and Computation Speed

The effect of post-training quantization on model size and computation speed was studied by comparing the models' size and computation time under different platforms. The first platform was a personal computer and the second platform was a Raspberry Pi Model B. The test was done under TensorFlow, a free and open-source software library for machine learning and artificial intelligence.

Raspberry Pi Model B is a series of small single-board computers. The Raspberry Pi used had the following hardware specifications: System on Chip (SoC) Broadcom BCM2837 1.2 GHz ARM Cortex-A53 Quad Core Processor (ARMv8 Family), Memory: 1 GB LPDDR2 running Debian operating system (OS). Raspberry Pi is a widely used embedded systems computer from applications such as robotics to weather monitoring. It is widely used because of its low cost, modularity, and open design.

The model was tested on personal computer. The personal computer CPU parameters were Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz, 32GB RAM, 2TB drive, running Ubuntu OS version 20.04. The model was run on the Jupyter notebook, a web-based interactive python computing platform.

Model	Model Size (KBytes)	Computation Speed (ms)	
		Personal Computer	Raspberry pi
MiniatureVQNet	137.5	97	-
Float16 MiniatureVQNet	58	8	36
Dynamic-Range MiniatureVQNet	62	6	25.2

Table 5 Post-training Quantized Models size and Computation Speed

The original trained model size was 137.5 Kilobytes (KB). Post-training model quantization reduced the model size. Float16 MiniatureVQNet is a Float16 quantized model of the original MiniatureVQNet, the model size was compressed from 137.5 *KB* to 58 *KB*. Dynamic-Range MiniatureVQNet is a dynamic range quantized model. The original MiniatureVQNet model was reduced from 137.5 *KB* to 62 *KB* after dynamic range quantization.

The model computation speed was measured based on the elapsed real time between inference invocation and termination in milliseconds (*ms*), the result is as show in Table 5. When the lightweight model was run on personal computer system, there was about tenfold improvement in speed. On personal computer the Float16 and dynamic range optimized model computation time were 8 *ms* and 6 *ms* respectively.

Running the model on Raspberry Pi, the Float16 and dynamic range optimized model computation time were 36 *ms* and 25.2 *ms* respectively. Post-training quantization improves model size and computation requirements, and makes the model suitable to run on embedded and resource constrained computer systems.

CHAPTER 6 PROPOSED MINIATUREVQNET MODEL EVALUATION

Ideally, the objective speech quality measure in VoIP applications should be able to estimate the quality independent of the type of speech distortions introduced by the VoIP system, regardless of whether they are network distortions, speech encoding–decoding, environmental noise, or the speech enhancement algorithm. This is highly challenging in IP audio applications as there are many factors, and the DNN training dataset must include at least all of these factors. In (Alkhawaldeh et al., 2019), the impact of network degradation features was studied independently on transmitted voice quality by using their MOS_LQO values and fixing the values of the remaining features.

The results showed that each feature independently has an effect and can be used as a feature in the training set. In evaluating the proposed MiniatureVQNet single-ended method, the accuracy of the MiniatureVQNet was examined on different environmental noise features and transmission network quality features. The acoustic features considered during evaluation were the noise type and signal–to–noise ratios, whereas the network factors examined were the packet loss, delay, and jitter. G722 audio codec was used to encode the speech signal for streaming.

Two variations of the proposed MiniatureVQNet model were evaluated, model trained on dataset with environmental noise only and the model trained on noise–network distorted speech dataset. The model trained on environmental noise only dataset is referred to as MiniatureVQNet–Noise throughout this work. Whereas, the model trained on noise-network distorted dataset is referred to as MiniatureVQNet–Noise–Network.

6.1 General Performance

First, we compared the performance of MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network with the ITU-T P.563 recommended non-intrusive speech quality evaluation method, on the test dataset which included clean, environmental noise and noise-network distorted features. Both MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network outperformed the P.563 in prediction accuracy. The mean squared error (MSE) of the models compared to the PESQ score for ITU-T P.563, MiniatureVQNet–Noise, and MiniatureVQNet–Noise–Network was 2.19, 0.34, and 0.21, respectively.

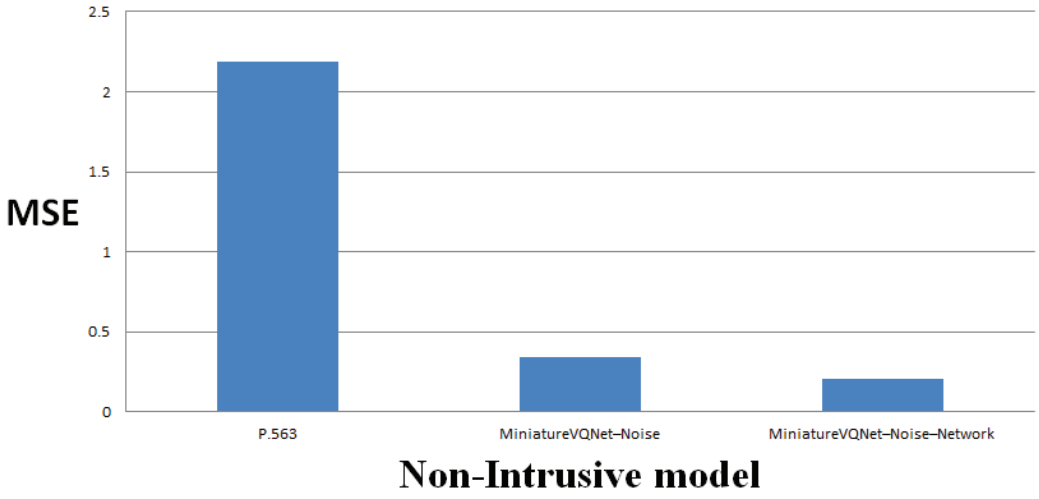

Figure 7 P.563, MiniatureVQNet–Noise, and MiniatureVQNet–Noise–Network Quality prediction comparison on clean, environmental noise and noise-network distorted features.

We expected P.563 to be outperformed by the proposed models. However, the difference in accuracy was out of our expectations. We were interested to know the cause of poor performance of P563 model. A further examination of the difference in performance of the P.563 and the MiniatureVQNet variants shows that, the difference is attributed to the failure of the P.563 model to accurately predict the speech quality when speech distortions are low, or the SNR is high. For high quality speech signals, the P.563 prediction is very poor, while for low quality speech signal the P.563 method accuracy is comparable to the MiniatureVQNet models, as shown in the Figure 8 and Figure 9 below.

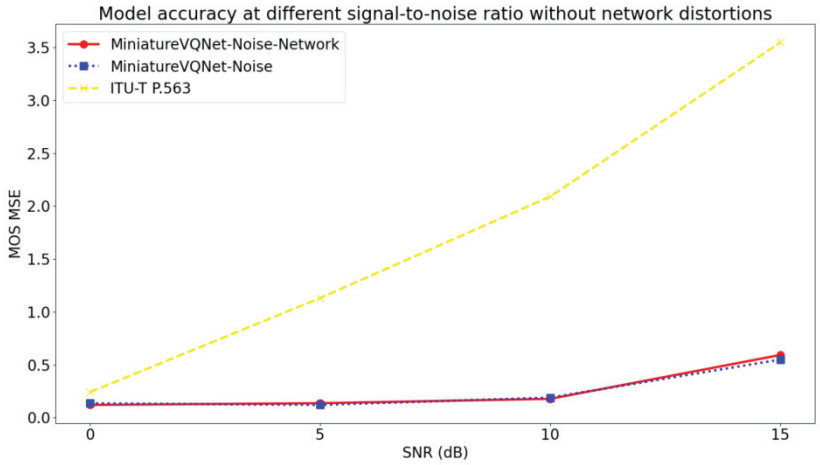

Figure 8 Model accuracy at different signal-to-noise ratios without network distortion

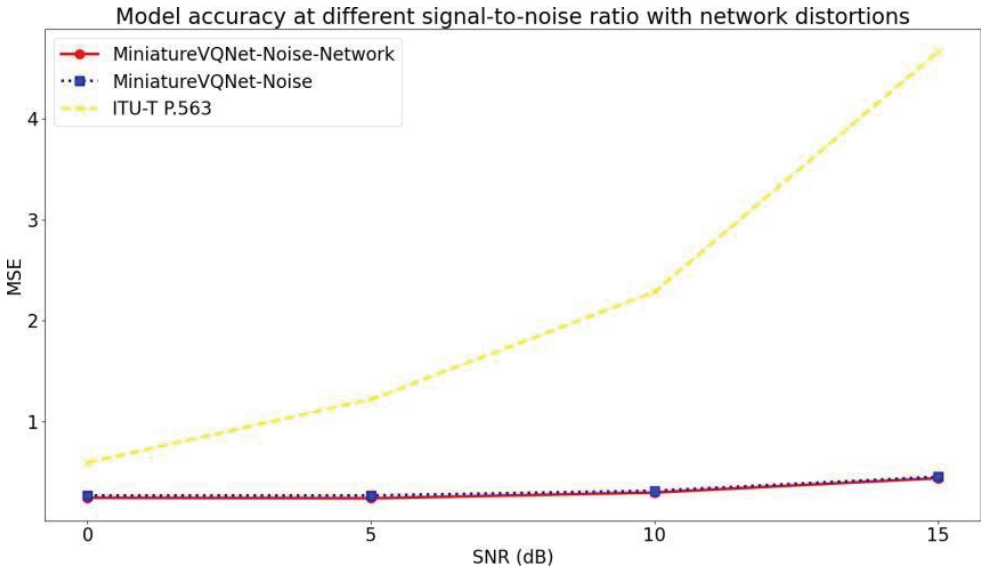

Figure 9 Model accuracy at different signal-to-noise ratios with network distortions

Figure 8, considers the effect of noise only, the P.563 MOS score MSE is almost equal to MiniatureVQNet model MOS score at low SNR. However, the error increases with the increase in SNR. Compared to Figure 9, even for low SNR, at 0 dB SNR the accuracy is different between the MiniatureVQNet model and P.563 method. DNN model performed better than the traditional P.563 model in both noise only, and noise-network condition.

6.2 The Effect of Noise and Network Distortion on Prediction Accuracy

While both the MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network models show a better performance than that of P.563, the MiniatureVQNet–Noise–Network, model trained on the noise–network speech dataset exhibits a higher accuracy than the MiniatureVQNet–Noise model, which is trained on noise only dataset.

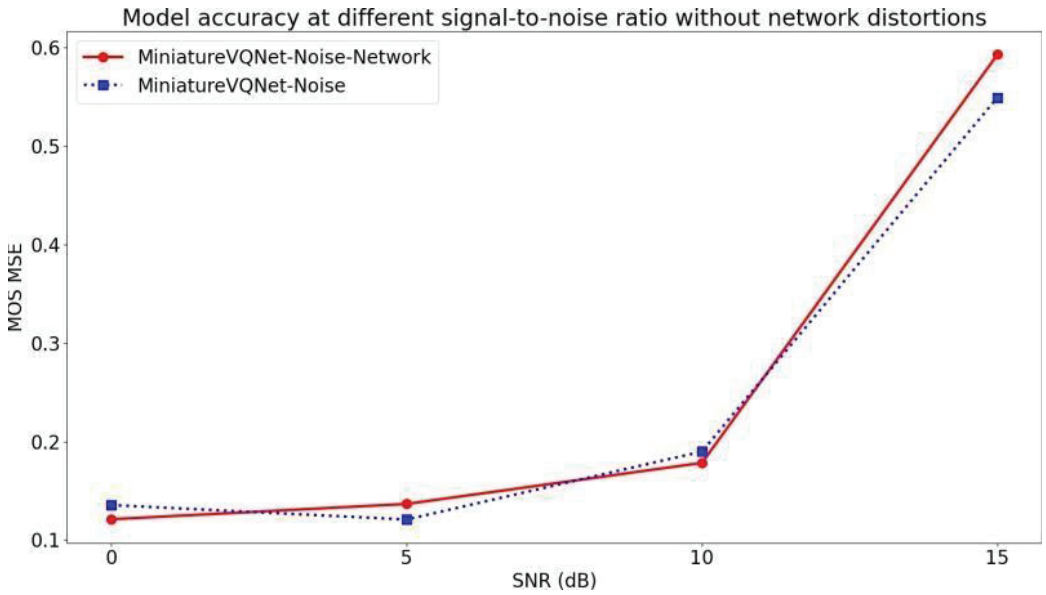

Figure 10 Comparison of MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network Models' accuracy at different signal–to–noise ratios on noise distorted speech data.

In the noise-only dataset, a dataset without network distortions (no packet loss, the delay is less than 3ms, no jitter), both models MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network exhibited the same performance. There is no significant difference between the model trained on the noise–network dataset and the model trained on the noisy-only dataset, as shown in Figure 10. This is interesting as the model trained on the noise–network dataset improves the general performance, without a decrease in performance on the noise-only dataset. The model trained on the noise–network–distorted dataset can learn network distortion features without a loss of knowledge on noise features. Thus, the proposed model is suitable for deployment in noise and noise–network conditions.

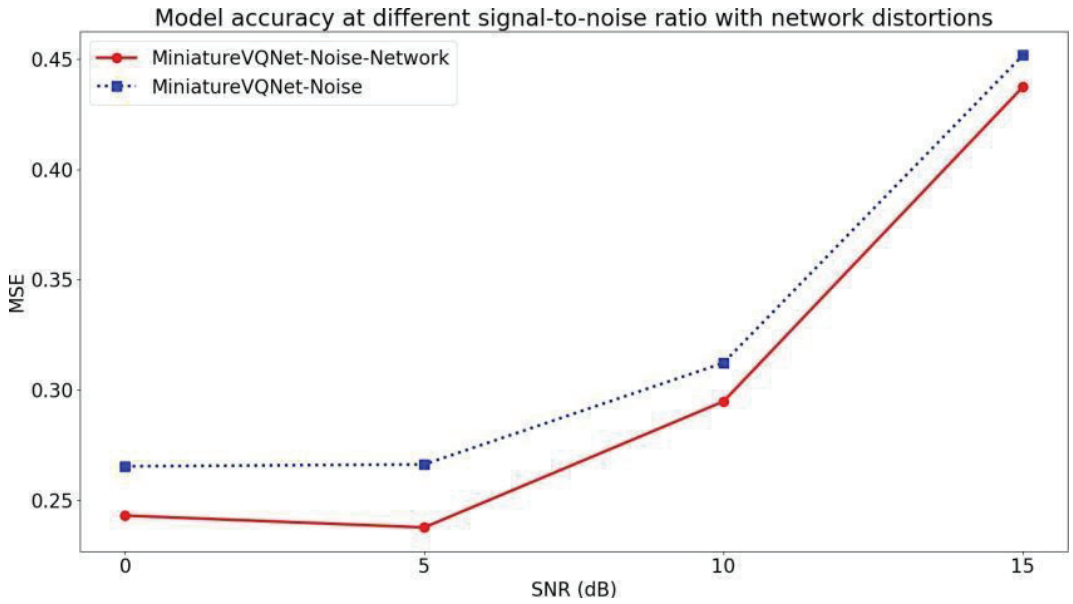

Figure 11 Comparison of MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network Model's accuracy at different signal-to-noise ratios on noise–network–distorted speech data.

Comparing performances on noise-network distorted speech dataset. The MSE of the MiniatureVQNet–Noise–Network model's accuracy is low compared to MiniatureVQNet–Noise as shown in the Figure 11. At a 0 dB signal–to–noise ratio, the MSE for MiniatureVQNet–Noise– Network and MiniatureVQNet–Noise models was 0.237 and 0.267, respectively. However, with the increase in SNR, both models' performances decreased. For example, at a SNR of 15 dB, the MSE of MiniatureVQNet–Noise was 0.452, whereas that of MiniatureVQNet–Noise–Network was 0.438. The performance decreases as the effect of network distortions is less at a high SNR than at a low SNR. Still, the noise–network–trained model outperforms the noisy dataset trained model at all SNRs.

In both cases, noise only speech dataset and noise-network distorted speech dataset MiniatureVQNet–Noise–Network exhibits higher performance compared to MiniatureVQNet–Noise model. Therefore, for single-ended speech quality monitoring in VoIP applications, model trained on noisy only datasets is outperformed by the model trained on noise–network speech datasets.

6.3 The Effect of Noise Type and Network Distortion on Prediction Accuracy

Different noise types have different characteristics. Therefore, we intend to understand the influence of various noise types at different SNR values on the performance of MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network models. Moreover, we intend to understand how network distortion impacts different noise types.

Noise type affects the model performance differently as shown in Figure 12 and Figure 13, where a comparison of street and train station noise is plotted. Train station noise constitutes noise from different sources, such as approaching trains, broadcasting loudspeakers, and conversation from nearby people. Street noise constitutes noise from passing cars, singing birds, and other sources.

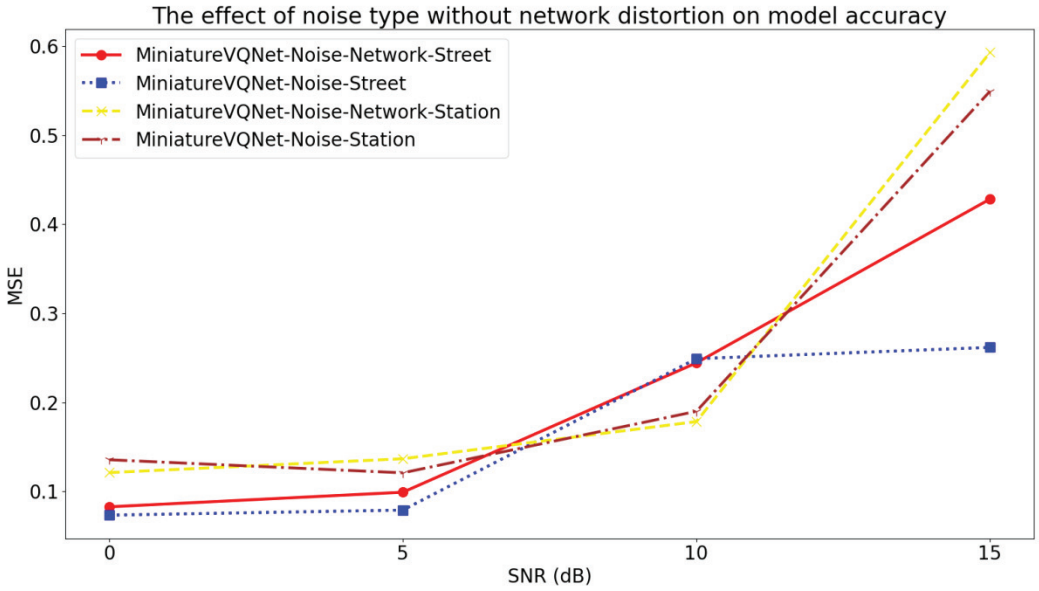

Figure 12 The effect of noise type without network distortions on prediction accuracy.

Figure 12 shows the MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network model performances on street and station noise. There is no significant performance difference for both street and station noise type for SNRs below 10 dB. When the speech SNR is greater than 10 dB, there is significance difference between different models performance on street noise, at high SNR station noise is more distorted than street noise. The MiniatureVQNet–Noise performs better than the MiniatureVQNet–Noise–Network model.

Generally, MiniatureVQNet–Noise–Network performs better than the MiniatureVQNet–Noise model, but care should be taken as there are cases where prediction performance on noise only speech dataset is degraded by training the model on noise-network distorted speech dataset.

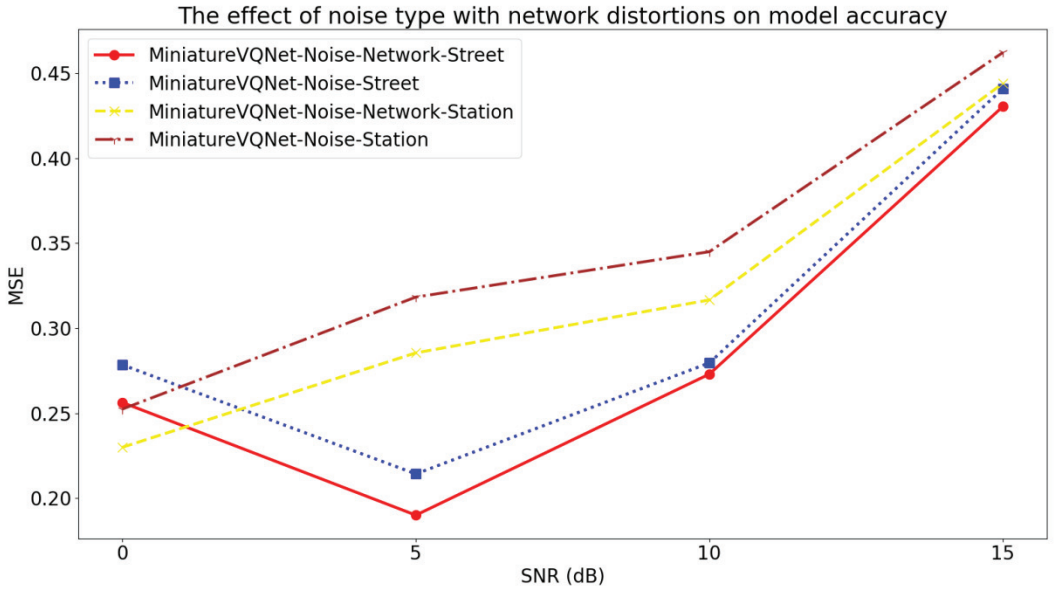

Figure 13 The effect of noise type with network distortions on prediction accuracy.

To investigate the mutual effect of network distortions and noise type on speech quality, the comparison of street and train station noise is shown in Figure 13. In comparison to noise only speech dataset, models' performance on noise–network distorted speech dataset depends on noise type and noise distortion (SNR).

Both models, MiniatureVQNet–Noise and MiniatureVQNet–Noise–Network performs better on station noise than on street noise. However for speech with SNRs less or equal to 0 dB and greater than or equal to 15 dB the performance difference based on noise type diminishes. Therefore, the effect of network distortion on noisy speech is more prominent for SNRs between 0 dB and 15 dB.

The DNN model prediction accuracy depends on the noise type. Nevertheless, training the model on the noise–network dataset improves the model accuracy on all SNRs and noise types. MiniatureVQNet–Noise–Network is more robust than MiniatureVQNet–Noise on all SNRs and noise types.

6.4 Effect of Jitter on Prediction Accuracy

Figure 14 shows the effect of jitter on the MiniatureVQNet–Noise–Network and Miniature VQNet–Noise model performance. We examined the effect of jitter at a 200 ms delay. The models showed significant performance differences at a jitter of 10% delay, which is 20 ms. At this jitter, the model trained on the noise-only dataset showed a 0.268 MSE, whereas that of the noise–network-trained model showed a 0.211 MSE.

In all other cases, the noise–network-trained model outperforms the noise–only–trained model, and the difference is large when the jitter is 10% of the delay. Training the model on the noise–network speech dataset improves the model's accuracy on different jitter distortions.

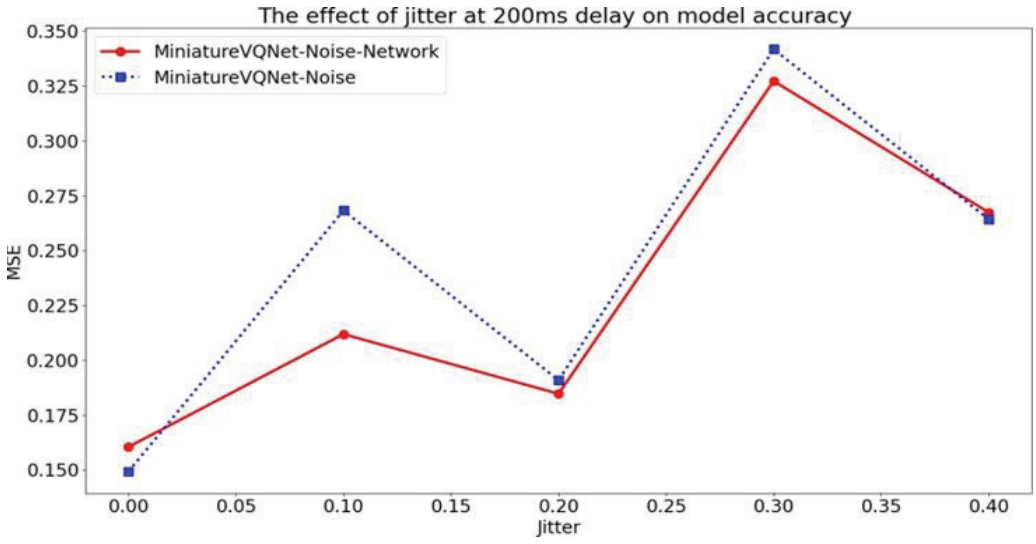

Figure 14 The effect of jitter at 200ms delay on model accuracy

6.5 Effect of Packet Loss on Prediction Accuracy

With the jitter, delay, and burst-loss kept constant, the effect of packet loss on the performance of the noise-trained model and noise–network-trained model was compared, as shown in Figure 15. The performance of the MiniatureVQNet–Noise–Network and MiniatureVQNet–Noise models decreases with the increase in the packet loss; the MSE

increases with an increase in the packet loss. Meanwhile, at all loss measurements, the MiniatureVQNet–Noise–Network model outperforms the MiniatureVQNet–Noise model. However, the performance difference is not very high, except for a 10% and 20% packet loss. Training the model on the noise–network dataset offers a better MOS prediction accuracy compared to models trained on the noisy dataset under all packet loss conditions.

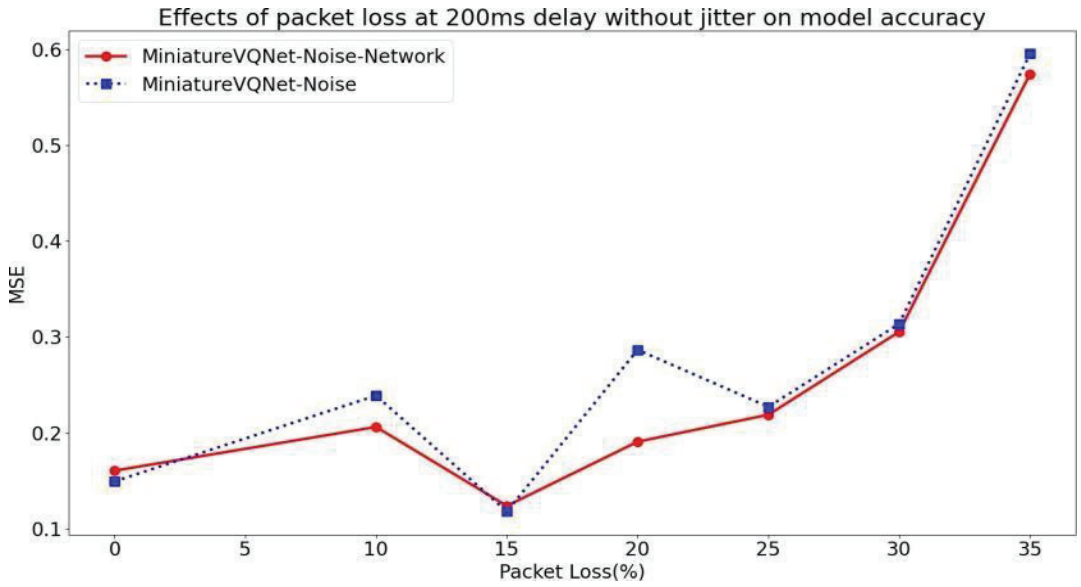

Figure 15 Effect of packet loss at 200ms delay without jitter on model accuracy

6.6 Post-Training Optimized Model Performance and Correlation

The MiniatureVQNet–Noise–Network was optimized after training to reduce the model size and execution time in low-powered devices. Two optimized models, Dynamic–Range MiniatureVQNet–Noise–Network and Float16 MiniatureVQNet–Noise–Network, were compared with the original MiniatureVQNet–Noise–Network.

Figure 16 shows MOS score distribution of the MiniatureVQNet dynamic rage (DNN DR) quantized model in comparison to P.563 method, using PESQ as a reference. MiniatureVQNet performs betters than P.563 in all MOS score distributions. MiniatureVQNet shows better prediction accuracy compared to P.563 on low MOS scores.

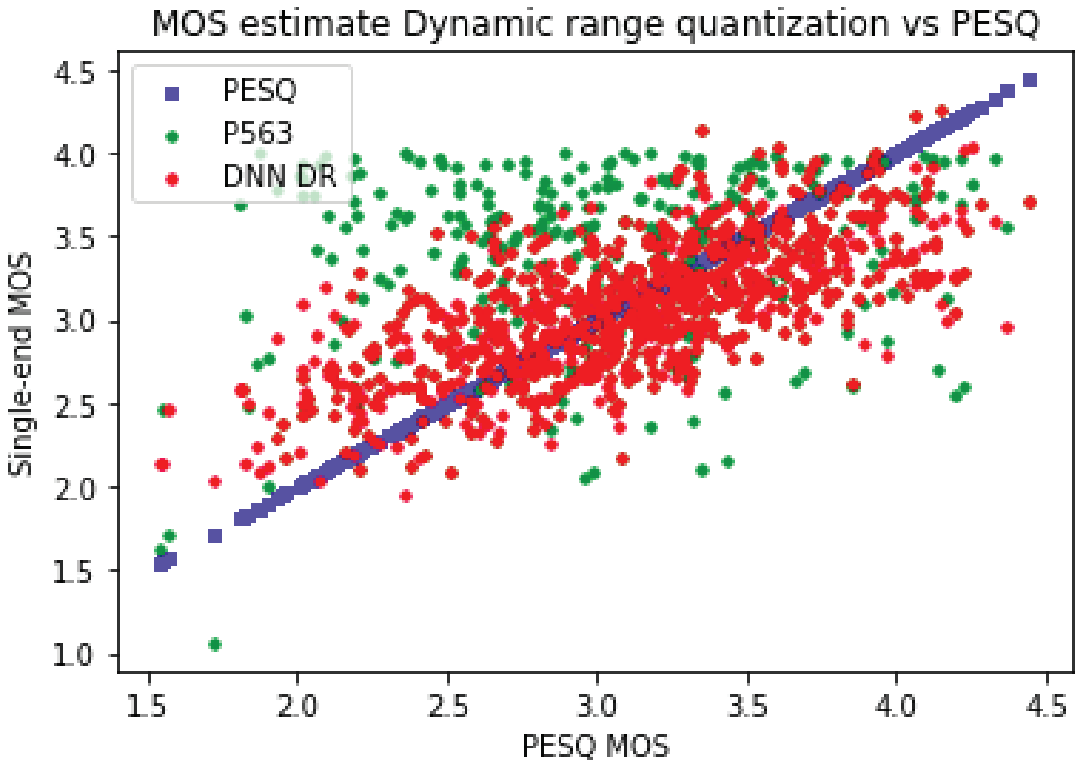

Figure 16 MOS comparison P563 and DNN Dynamic range quantized model

Table 6 shows the Pearson correlation coefficient of MiniatureVQNet–Noise–Network, Dynamic–Range Miniature VQNet–Noise–Network, and Float16 MiniatureVQNet–Noise–Network with respect to PESQ. For optimized models, there is a slight decrease in the Pearson correlation for Dynamic–Range MiniatureVQNet–Noise–Network compared to the original model, but this difference is not significant.

Model	Correlation	MSE
MiniatureVQNet–Noise–Network	0.691	0.194
Float16 MiniatureVQNet–Noise–Network	0.690	0.194
Dynamic-Range MiniatureVQNet–Noise–Network	0.670	0.254

Table 6 Correlation and Mean Squared Error.

The MSE of Dynamic–range MiniatureVQNet–Noise–Network was 0.254, which is high compared to MiniatureVQNet–Noise–Network and Float16 MiniatureVQNet–Noise–Network, which were each 0.194. The results show that model compression can be achieved without a loss of model accuracy. Therefore, the proposed model's weights representation can be reduced from float64/float32 to a low number of bits (float16) without a loss of accuracy. Hence, the quantized model can be used in a resource-constrained environment without a loss of accuracy.

6.7 Raw Sample Data of Speech Quality MOS Score for Different Models on Noise-Network Distorted Speech Dataset

The preceding performance evaluation graphs above show the general behavior of the proposed speech quality prediction model on clean speech and noise-network distorted speech dataset. Table 7 shows the result for an individual recorded speech utterance file at different network and environmental noise parameters.

Text	her purse was full of useless trash				
Distortion parameters	Clean	Street Noise SNR 5 dB	Station Noise SNR 15 dB	Station Noise SNR 15 dB + 200ms delay+10% packet loss	Street Noise SNR 15 dB + 200ms delay+30% packet loss
ITU-T PESQ (End to End)	**4.5**	**1.45**	**2.53**	**2.53**	**1.78**
ITU-T P.563	4.5	2.08	4.5	4.5	4.23
MiniatureVQNet Noise speech	4.4	1.69	1.65	1.68	1.67
MiniatureVQNet Noise-Network speech	**4.5**	**1.60**	**1.78**	**1.76**	**1.73**

Table 7 Raw Sample Data of Speech Quality MOS Score for Different Models on Noise-Network Distorted Speech Dataset

CHAPTER 7 IMPROVED AUTOMATIC SPEECH RECOGNITION ON NOISE–NETWORK DISTORTED SPEECH DATASET

Different ASR engines have been proposed. To evaluate the performance of ASR systems on the noise-network distorted database, we used DeepSpeech version 0.9.3 (Mozilla, 2020). DeepSpeech is an open-source speech-to-text engine that uses a model trained by machine-learning techniques based on Baidu's study (Hannun et al., 2014). This technique does not require hand-designed features to model background noise, reverberation, or phoneme dictionary. Instead, it relies on large amounts of varied data for training.

The aforementioned features render DeepSpeech Engine the best candidate for evaluating the performance on noise-network distorted speech. Evaluation was carried on both models, the DeepSpeech CSM and NNSM.

7.1 ASR Performance Metrics

An automatic ASR performance measurement is necessary for the rapid system development and the performance comparison of different ASR systems. Researchers generally report the performance of ASR systems using the Word Error Rate (WER) metric. WER is defined as the ratio of the total number of errors (substitution, deletion, and omissions) in the transcription output to the number of words in the speech signal input to the ASR system, given by the equation below.

$$WER = \frac{S + D + I}{N}$$

Equation 2 ASR Word Error Rate Function

,where S is the number of erroneous word substitutions, D is the number of word deletions, I is the number of insertions of false words in the ASR output, and N is the number of words actually spoken in speech input to the ASR system.

The WER does not reflect human judgment, such as the relative importance of certain words for the meaning of the message. Therefore, more intuitively appealing measures for

ASR – the match error rate (MER) and word information lost metric (WIL) – were also used (Morris et al., 2004). MER is the probability of a given match being incorrect, obtained by simply dividing the WER by its maximum possible value. Let H, S, D, and I denote the total number of word hits, substitutions, deletions, and insertions, respectively.

$$MER = \frac{S + D + I}{H + S + D + I}$$

Equation 3 ASR March Error Rate Function

The WIL metric is the difference between 100% word preservation and percentage on output words preserved. Where $H > S + D + I$ the word information preserved (WIP) is given by:

$$WIP = \frac{H}{N}$$

Equation 4 ASR Word Information Preserved Function

Then, WIL is derived from WIP, which is given by the equation below.

$$WIL = 1 - WIP$$

Equation 5 ASR Word Information Lost Function

The DeepSpeech model performance was tested using both clean speech and noise-network speech, before and after fine-tuning with the noise-network speech dataset. The metrics used for comparison were the WER, WIL, and MER. In the next section, the results of the comparison between the performances of the two models are presented.

7.2 Dataset

The datasets used for the training and testing of deep-learning-based ASR systems has evolved from clean-read speech, spontaneous-speech, large dataset size speech corpus, artificially added environmental noise, speech recorded in domestic environments, and speech transmitted through cellular networks. The proposed dataset aims to build on the existing datasets with the addition of distortions induced by network conditions and encoding-decoding schemes on speech transmitted through the VoIP system. This dataset is built on clean speech, which is then distorted by noise at different signal-to-noise ratios (SNR), then transmitted at different network quality of service (QoS) parameters to generate noise-network distorted speech dataset.

The noise-network distorted speech dataset was divided into a training dataset and a testing dataset. The testing data was selected from the total sample of noise-network distorted dataset using stratified random sampling. The total sample was divided into groups of sentences with similar attributes, as shown in Table 4. In each similar attribute group, there were 30 sentences, which were further divided into utterances of male and female speakers. Then, three samples were selected from each group. Finally, through the stratified random sampling, 20% of the total sample was set for testing, while the remaining 80% was used for training. This testing dataset sampling method was selected to ensure that different sample attributes were equally represented in training and testing datasets. One hundred and fifty four hours of speech were used for training and 38.5 hours of speech were used for testing.

7.3 ASR Pre-Trained Model and Fine-Tuning Process

The pre-trained model of DeepSpeech was trained on Fisher, LibriSpeech, Switchboard, Common Voice English, and WAMU radio-shows databases. The acoustic models were trained on American English with synthetic noise augmentation, and the model achieved a 7.06% word error rate on the LibriSpeech clean test corpus. The performance of the pre-trained DeepSpeech CSM on noise-network distorted speech dataset was analyzed, and then the model was optimized through transfer learning.

Transfer learning transfers the knowledge gained when solving one problem and applies it to a different problem in a related domain. Fine-tuning is a transfer learning technique that starts with a pre-trained model on the source task and trains it further on the target task.

Fine-tuning is a common technique in computer vision tasks (Kornblith et al., 2019). In ASR, fine-tuning was successfully applied in low resource languages, where models trained to recognize speech in rich resource languages were then transferred to low resource languages [(Huang et al., 2013), (Kermanshahi et al., 2021), (Shi et al., 2019)]. To the best of our knowledge, this is the first case the transfer of knowledge gained by a model in speech–to–text conversion of clean speech is applied on speech–to–text conversion of noise-network distorted speech.

The noise-network dataset contained the same alphabet set as the dataset used to train the DeepSpeech model. Therefore, the released DeepSpeech model output layer matches noise-network data, and there were no need for a different classifier in our experiment. We fine-tuned the entire model graph with the noise-network dataset without adding new layers. In this experiment, the model parameters and architecture were equal to those of the released DeepSpeech model with the training dataset as the only difference.

Hence, transfer learning on noise-network dataset was evaluated as the sole factor to the ASR performance improvement. The training system environment had the following hardware and software specifications. Hardware specifications were: - Processor: Intel® Core™ i7-9750H CPU @ 2.60GHz × 12, Graphics: NVIDIA Corporation TU117M [GeForce GTX, 1650 Mobile / Max-Q] / GeForce GTX 1650/PCIe/SSE2, Memory: 31.2GB, Disk capacity: 1.3TB. While the OS platform used was Ubuntu 20.04.2 LTS, 64-bit, and GNOME Version:3.36.8 with Windowing System: X11.

The open-source TensorFlow framework was used to build the model and train the network. The model network architecture was the same as that of DeepSpeech. The network was trained in six stages of 10 epochs each, and a generalization evaluation was performed at each stage. The training, testing, and validation used the following parameters: The training batch size, test batch size, and validation batch size were 112, as the dataset was in the multiples of 112. This is different from the original DeepSpeech model, which used a batch size of 128 for training, testing, and validation. As in the original model, a training learning rate of 0.0001 and dropout rate of 0.4 were used. For each stage, the generalization performance of the network was monitored using a subset of the Mozilla Common Voice Corpus 1 English dataset. This speech dataset is referred to as the clean speech in the experiment result presentation.

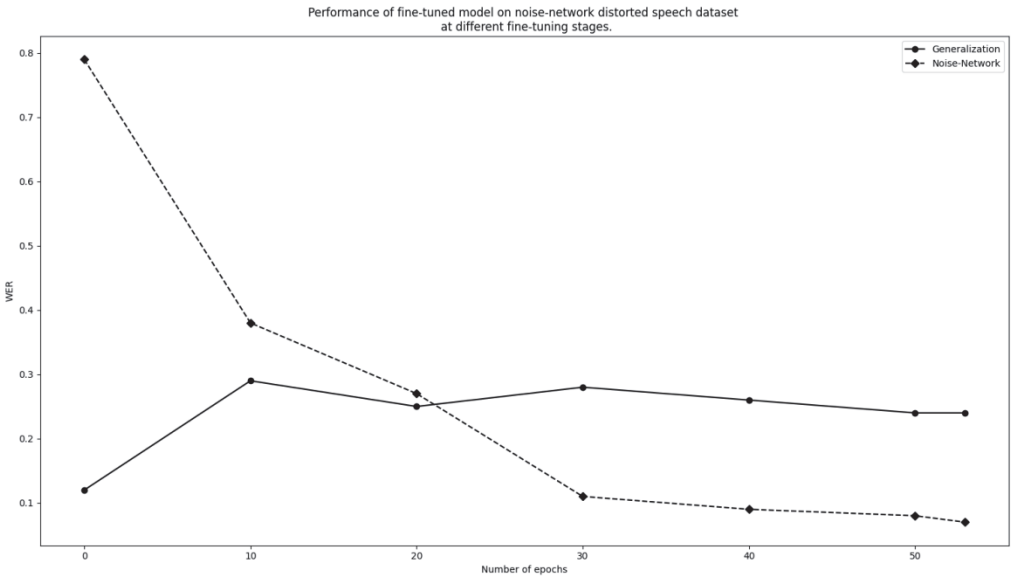

Figure 17 Transfer learning accuracy and generalization performance

Fine-tuning ASR model on noise-network distorted speech dataset improved the performance. However, the model's generalization performance slightly decreases. WER before fine-tuning was 0.12 on Mozilla Common Voice Corpus 1 English dataset, by fine-tuning the model on noise-network distorted dataset the WER increased to 0.26 and then decreased to 0.23 as shown in Figure 17.

Fine-tuning the model leads to decrease in performance on the original dataset, but the decrease is not proportional to the increase in performance on fine-tuning dataset. Comparing the model performance on noise-network distorted speech dataset and that of Mozilla Common Voice Corpus 1 English dataset. Before fine-tuning the models' WER on noise-network distorted dataset was 0.79 whereas after fine-tuning the performance was 0.07. The models performance decrease reaches a point where further fine-tuning does not affect the generalization performance.

7.4 Experiment Results

Fine-tuning a CSM on a noise-network distorted speech dataset improved model performance on noise-network distorted speech. However, the ASR model fine-tuned on noise-network speech undergoes a slight degradation on the generalization performance on clean speech compared to the ASR model trained on clean speech.

The WER of the DeepSpeech model on the clean speech dataset was 0.12 and 0.24 before and after fine-tuning, respectively. However, the model performance on the noise-network distorted speech dataset improved significantly from 0.79 before fine-tuning to 0.07 after fine-tuning. The ASR performance on the noise-network distorted speech improved at the expense of generalization performance, but the degradation was less when compared to the improved robustness. **Robustness** in this study means the ability for the model's performance to withstand changes in input speech degraded by noise-network distortions. The models' accuracy is not to be affected by speech degradation.

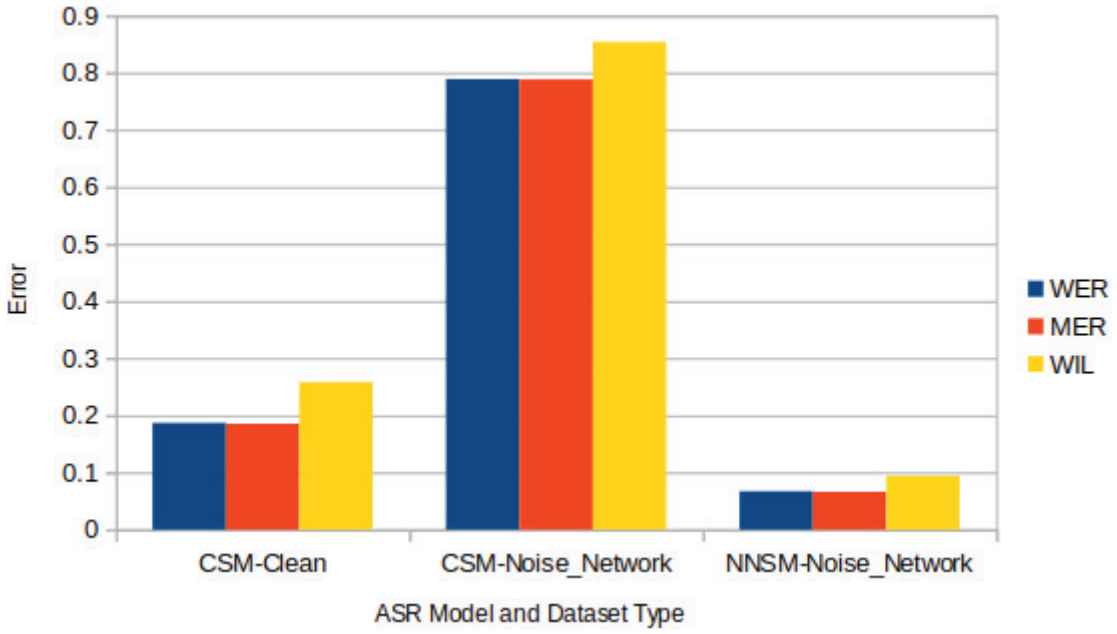

Figure 18 ASR model's performance before and after fine-tuning on noise-network distorted speech dataset

The performance of CSM on clean and noise-network-distorted speech datasets was compared with that of the NNSM on the noise-network-distorted speech dataset. Generally, noise-network distortions resulted in equal degradation on WER, MER, and WIL, with WER and MER increasing from 0.19 to 0.79 and the WIL rate increasing from 0.24 to 0.85. The fine-tuning improved the model performance on WER and MER from 0.79 to 0.07, while WIL decreased from 0.85 to 0.09.

7.4.1 Isolated Effect of Noise and Network Distortion on WER

To examine the individual effect of noise distortion and network distortion on noisy speech data, the noise distorted speech data without any network distortion was used. Then, the general network distortion effect for each SNR was observed. The performance of the two models, the CSM and NNSM are shown in Figure 19.

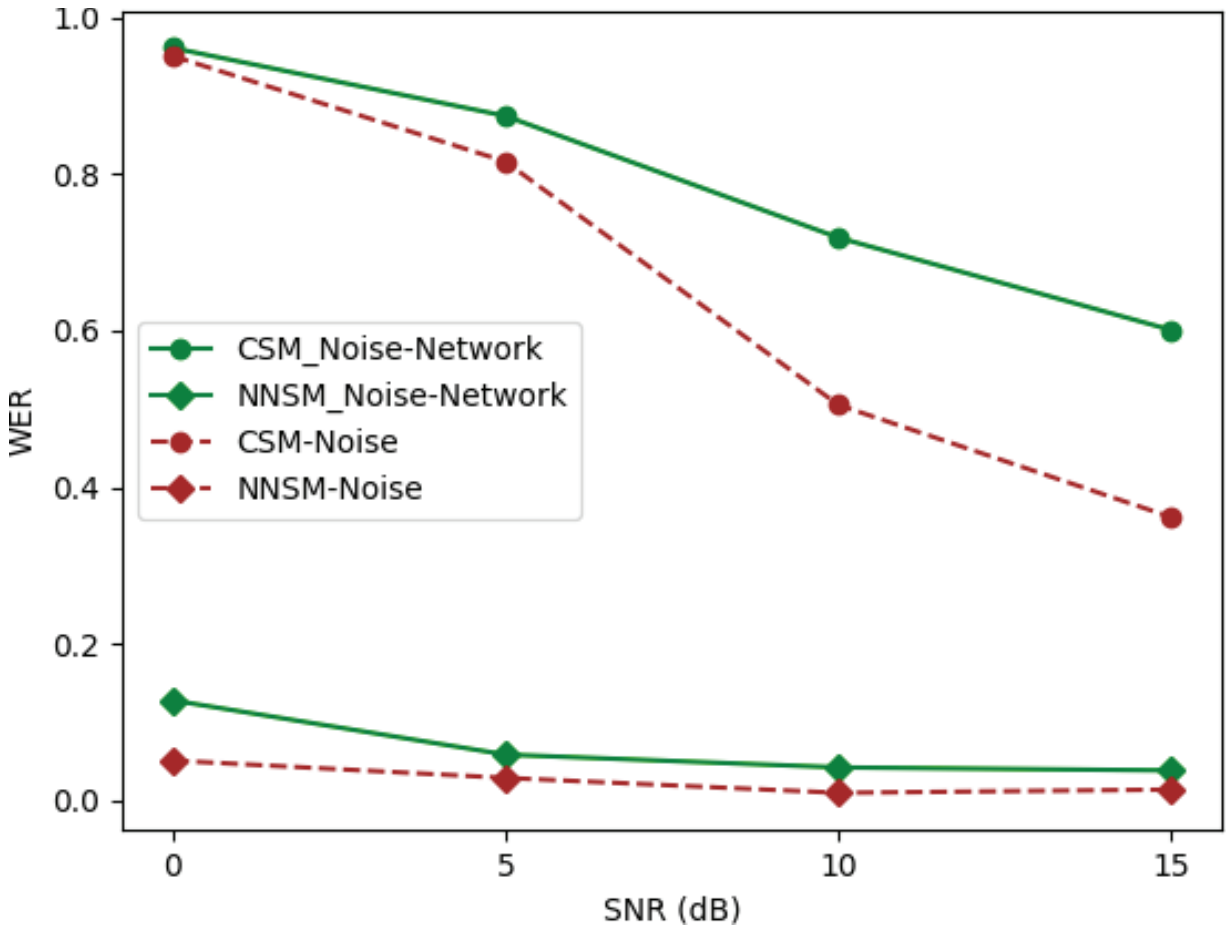

Figure 19 Effects of SNR and network distortion on WER.

As expected, WER decreased with an increase in the SNR. However, it was noticed that the network distortion effects were high on high SNRs and did not cause significant differences to speech signals with low SNR, which were already highly distorted by noise. The fine-tuned model exhibited the same performance for SNR greater than 5 dB. However, the

performance decreased significantly for the SNR less than 5 dB. Therefore, NNSM has improved robustness for speech signals with SNR greater than 5 dB, independent of network distortions.

7.4.2 Effect of Noise Type and Network Distortion on WER

Different noise types have different characteristics. We intend to understand the influence of various noise types at different SNR values on the performance of CSM, and NNSM. Moreover, we intend to understand the manner in which the network distortion impacts different noise types. Noise type affects WER, WIL, and MER differently, as shown in Figure 20 – by a comparison of street noise and train station noise.

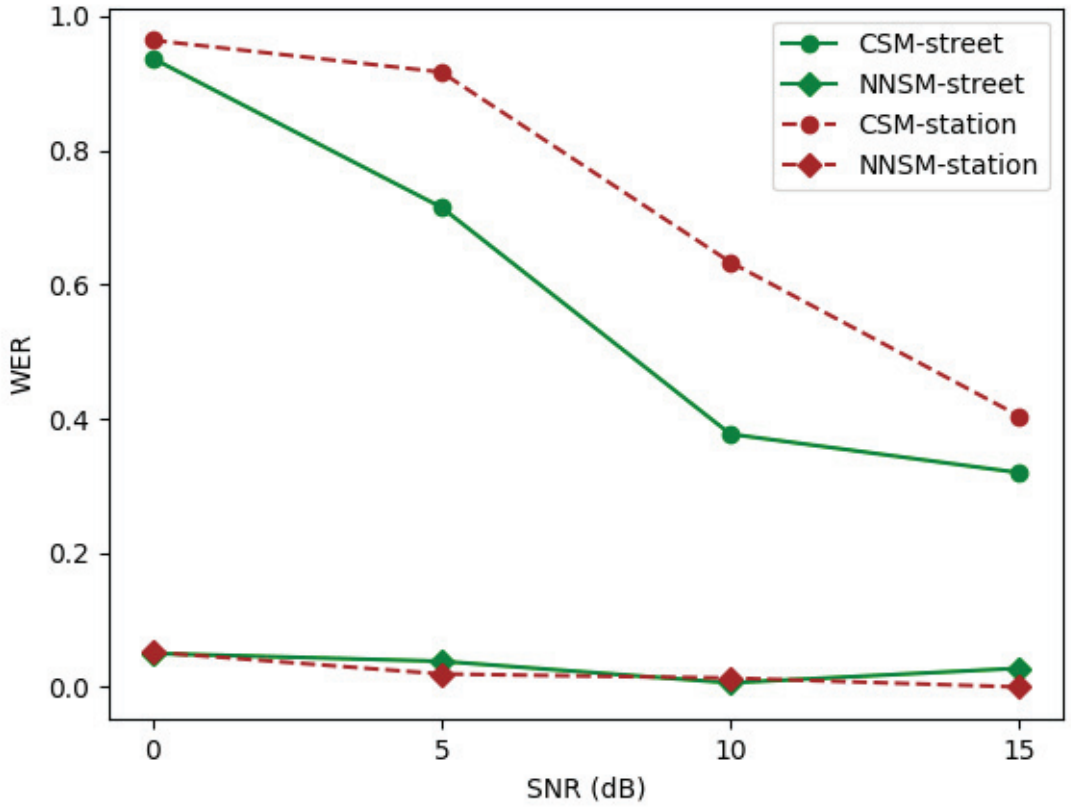

Figure 20 Effects of street noise and station noise with network distortion on WER.

Train station noise constitutes the noise from different sources, such as approaching trains, public addressing speakers, and nearby conversations. Street noise constitutes the noise

from passing cars, singing birds, and other sources. The CSM performance on station noise was lower than that of the CSM on the street noise for all SNR values. Moreover, the robustness of the fine-tuned network using noise-network distorted speech is evident, as the performance of NNSM is not affected by noise type.

When the noise-distorted speech was further distorted by network transmission errors, there was no difference in the performance of CSM and NNSM on different noise types as shown in Figure 21.

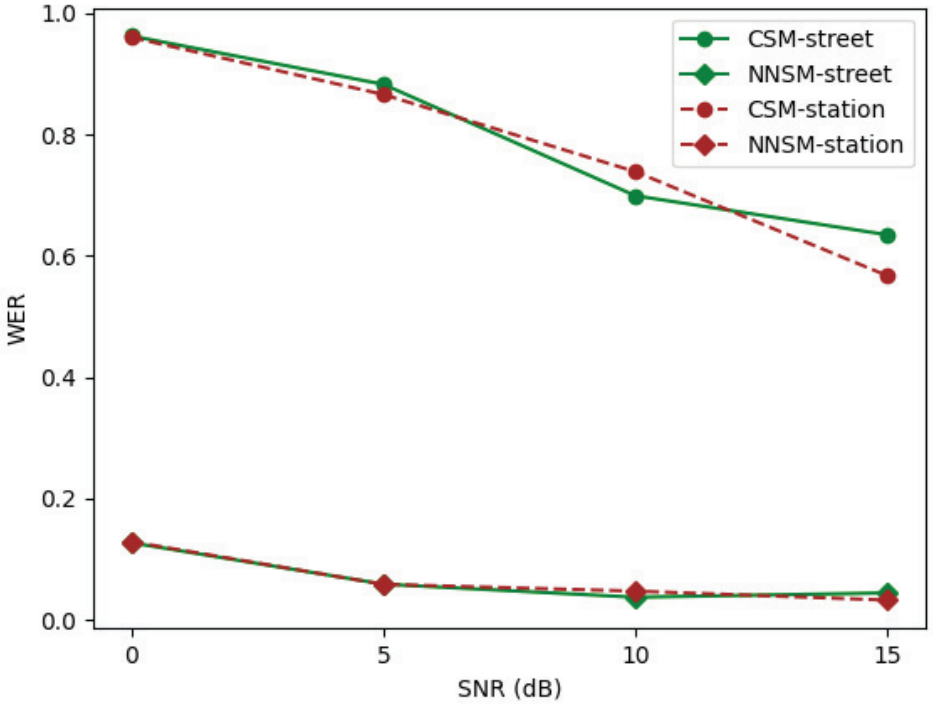

Figure 21 Effect of street noise and station noise with network distortion on WER.

The network distortion on noisy speech masks the noise effects on WER. The WER increased with a decrease in the SNR of the speech for the CSM, but for the NNSM, the performance was the same for SNR values greater than 5 dB. The NNSM performance on noisy speech and noise-network distorted speech data deteriorates for the SNR values of less than 5 dB. The NNSM can learn the effect of noise-network distortions when the SNR is greater than 5 dB. Hence, the NNSM is more robust than the CSM.

7.4.3 Effect of Jitter on WER, MER, and WIL

Figure 22 shows the effect of jitter on WER, MER, and WIL. If all network parameters are constant and the jitter is less than 0.2 of delay, there is a constant effect on WER, MER, and WIL. However, with a jitter greater than 0.2 of delay, WER, MER, and WIL begin to increase proportionally to the jitter.

For the CSM, WIL is greater than MER and WER. However, for the NNSM, WER, MER, and WIL are equal when the jitter is less than 0.2 of delay, with the WIL higher than the WER and MER when the jitter is greater than 0.2 of delay.

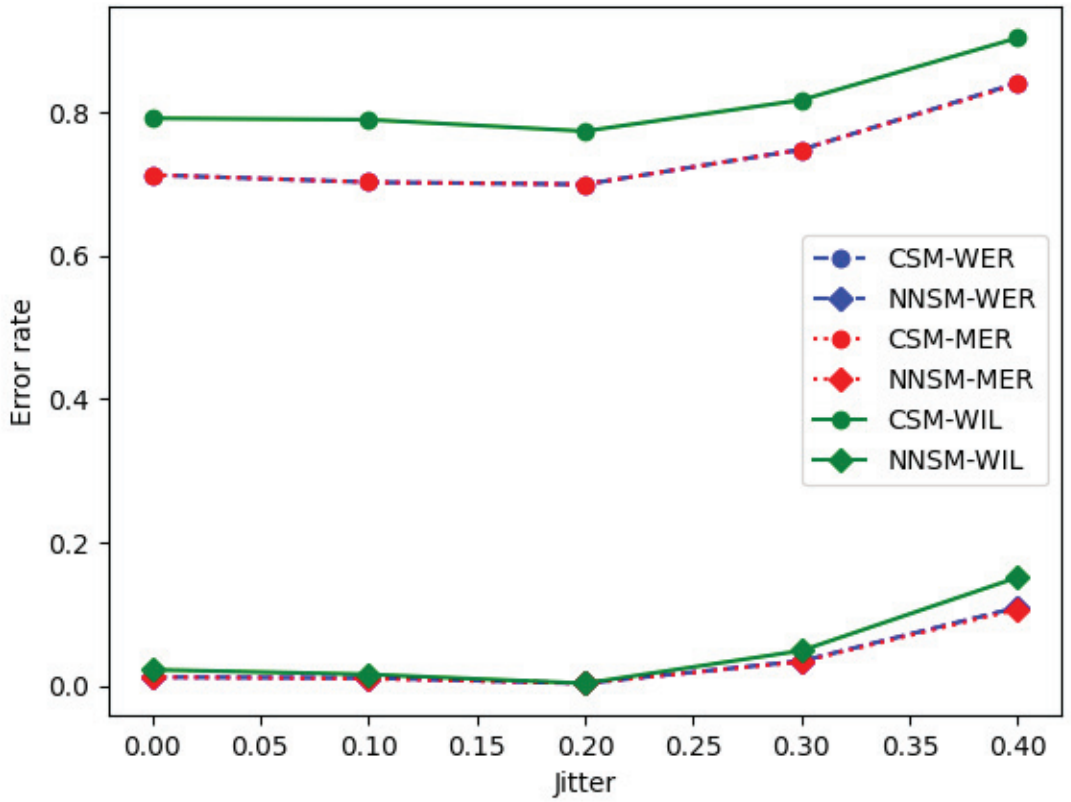

Figure 22 Effect of jitter on WER, MER, and WIL, for a delay of 200ms with packet loss of 0%.

7.4.4 Effect of Packet Loss on WER, MER, and WIL

With the jitter, delay, and burst-loss kept constant, the effect of packet loss on the clean-speech-trained model is constant for loss less than 10%. However, for a packet loss greater

than 10%, the WER, MER, and WIL increased proportionally to the packet loss. By contrast, for the noise-network-trained model, the effect of an increasing packet loss begins to be seen when the loss is greater than 15%. When the packet loss is greater than 15%, the WIL error rate increase is greater than that of WER and MER, as shown in Figure 23.

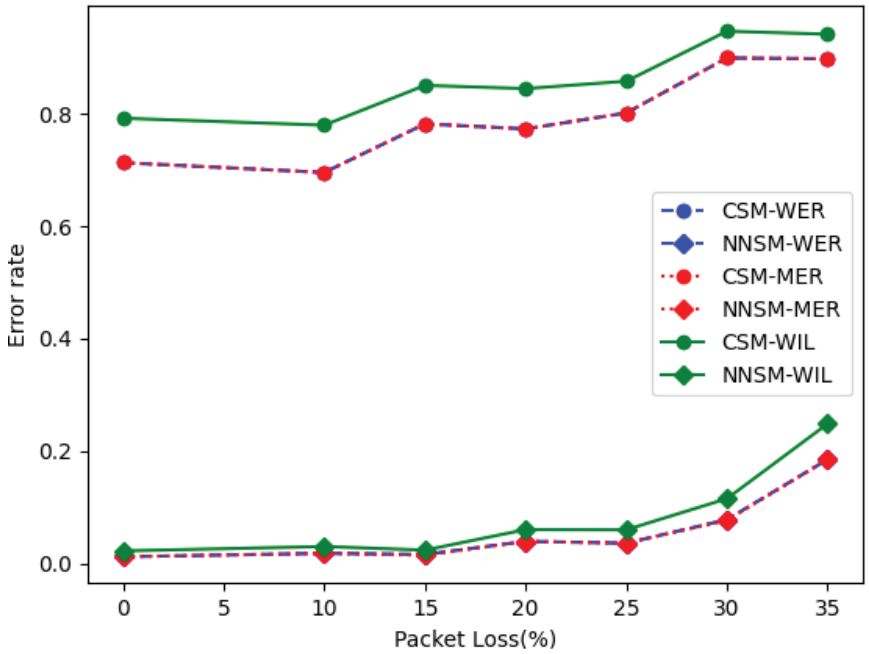

Figure 23 Effects of packet loss on WER, MER, and WIL, for a constant delay of 200ms without jitter.

7.4.5 Combined Effect of SNR and Packet Loss on WER

The combined effect of SNR and packet loss shows that both SNR and packet loss contribute significantly to the decreased performance of the clean-speech-trained ASR as shown in Figure 24. The WER of clean-speech-trained ASR model increases with an increase in packet loss for all SNRs, whereas the WER increases with the decrease of SNR. However, for an SNR of 0 dB, the effect on WER is dominated by SNR rather than packet loss.

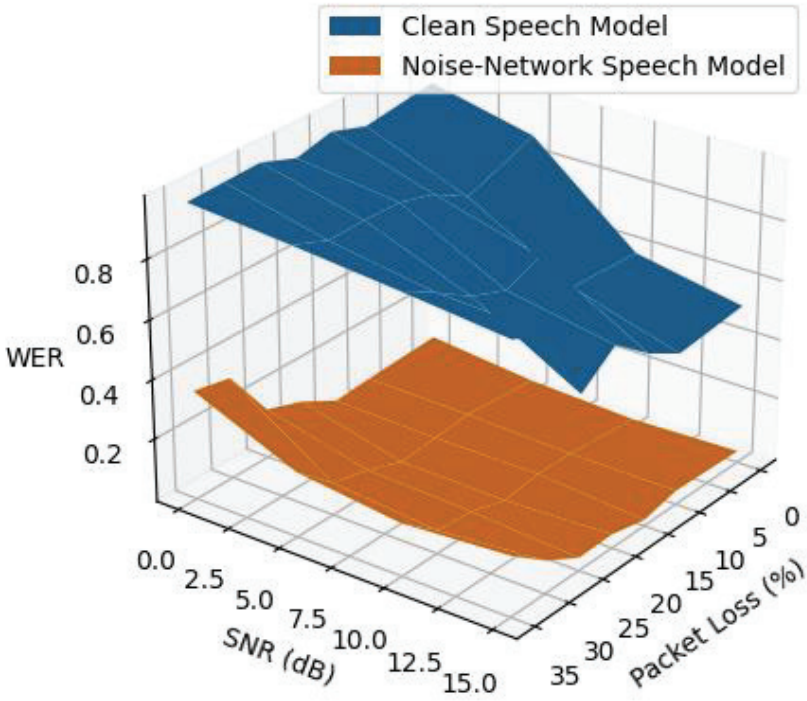

Figure 24 Effects of SNR and packet loss on WER, for a constant delay of 2 ms without jitter.

The NNSM yields significantly improved performance compared to that of the CSM. Furthermore, the NNSM shows better robustness compared to the CSM. A change in packet loss and SNR generates a small change in accuracy of the NNSM compared to that of the CSM. The NNSM performance can withstand the packet loss of less than 15% and the SNR values greater than 5 dB without loss of accuracy.

7.4.6 Combined Effect of SNR and Jitter on WER

An examination of the effect of jitter and SNR on ASR performance shows that the effect of SNR was significant compared to that of jitter, as shown on Figure 25. However, when the ASR is trained using noise-network speech distorted dataset, the robustness of the ASR increases. The effect of SNR and jitter is observed for the SNR less than 5 dB and the jitter greater than 30% of the delay, and the SNR effect does not dominate the jitter effect on the noise-network-trained ASR system.

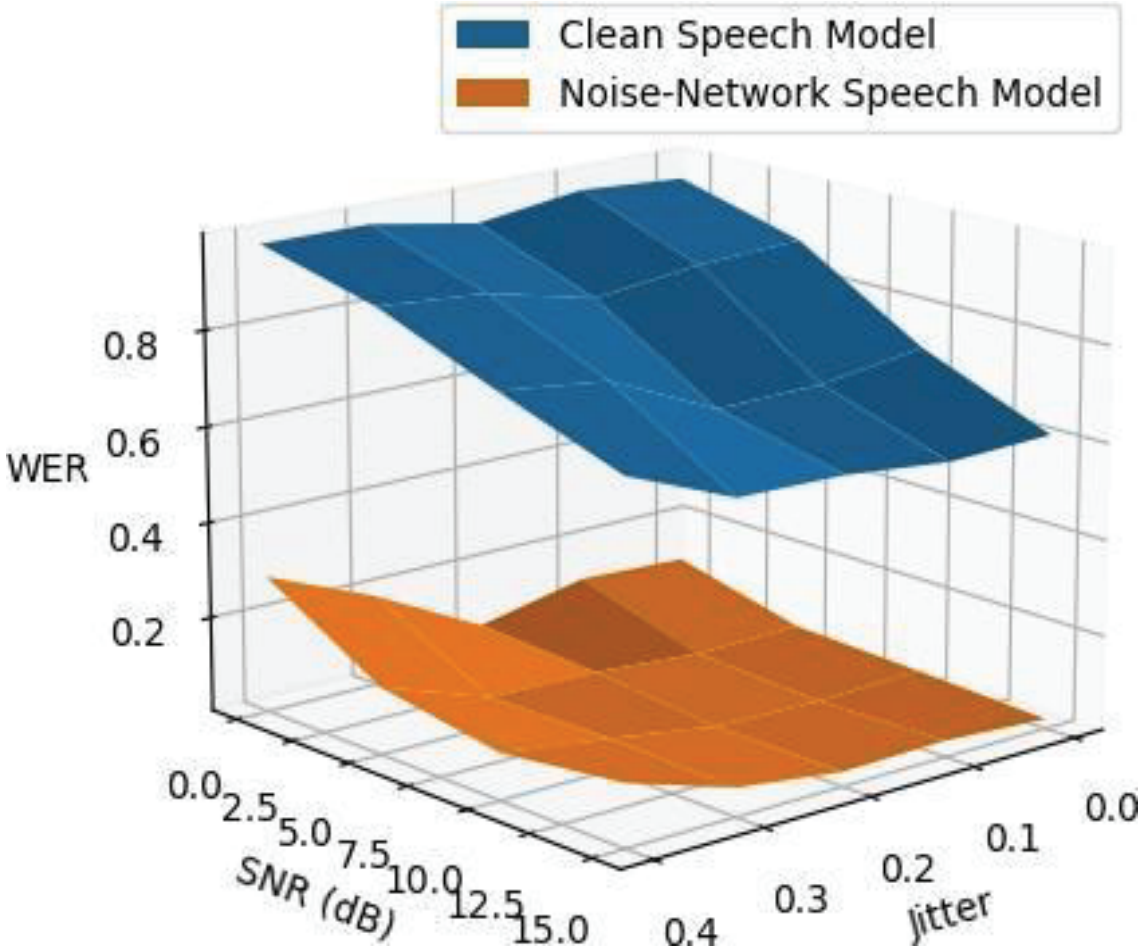

Figure 25 Effect of SNR and jitter on WER, for a delay of 200ms with packet loss of 0%.

7.4.7 Combined Effect of Jitter and Packet Loss on WER, MER, and WIL

The combined effect of jitter and packet loss has an increased impact on the performance of ASR for WER, MER, and WIL. Figure 26 shows the effect of jitter and packet loss on WER. Training ASR on noise-network distorted speech minimizes the WER. However, the improvement starts to decrease when the jitter is higher than 0.3 of delay and the packet loss is greater than 15%. The NNSM can learn the patterns for jitter and packet loss better than the CSM models.

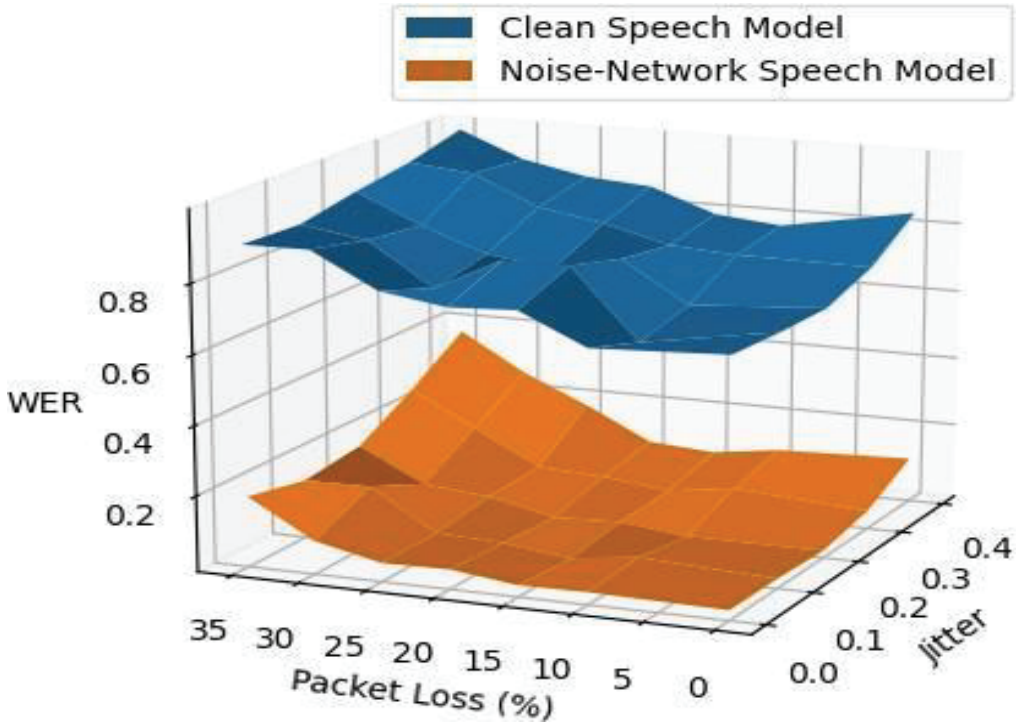

Figure 26 Effects of packet loss and jitter on WER, for at 200ms delay.

7.4.8 Raw Sample Data of ASR Evaluation WER Score for Different Models on Noise-Network Distorted Speech Dataset

The preceding performance evaluation graphs above show the general behavior of speech recognition system on clean speech and noise-network distorted speech dataset. Table 8 shows the result for an individual recorded speech utterance file.

Text	her purse was full of useless trash				
Distortion parameters	Clean	Street Noise SNR 5 dB	Station Noise SNR 15 dB	Station Noise SNR 15 dB + 200ms delay+10% packet loss	Street Noise SNR 15 dB + 200ms delay+30% packet loss
Clean Speech Model (WER)	0	0.833	0.167	0.333	0.833
Noise-Network distorted speech model (WER)	0	0	0.167	0	0.267

Table 8 Raw Sample Data of ASR WER Score for Different Models on Noise-Network Distorted Speech Dataset

Results show that there were no performance difference between the clean speech and noise-network distorted speech trained models on clean speech dataset. However, at low SNR (5 dB), and high network distortions noise-network distorted speech dataset model performance was better than the clean speech trained model.

CHAPTER 8 CONCLUSION

This work proposes the MiniatureVQNet, a single-ended speech quality evaluation method for VoIP audio applications, based on a lightweight deep neural network (DNN) model trained on an environmental noise and network-distorted speech dataset. The proposed model can predict audio quality independent of the source of degradation, whether noise or network, and is light enough to run in embedded systems. The proposed MiniatureVQNet model outperforms the traditional P.563 method in accuracy on all tested network conditions and environmental noise parameters. Furthermore, the proposed model is compact and can easily run on various low-resource computing platforms.

Training on a noise-network distorted speech dataset improves the DNN based speech quality prediction models' accuracy in all VoIP environment distortions compared to training the model on a noise-only dataset. The noise–network dataset captures speech degradation from environmental noise and also transmission-induced degradations. In all considered environmental noise and transmission error factors, the MiniatureVQNet trained on the noise–network–degraded dataset outperformed the model trained on the noise-only dataset. Hence, the MiniatureVQNet model can learn new features without degradation in performance. Examining the model performance on different noise types, SNR, jitter, packet loss, and delay provides the model's weak points, where the model performance is low. Care should be taken when deploying single-ended speech quality prediction models as the stated performance may not be applicable in all conditions.

Recently, music streaming, BGM streaming, Internet radios, and IP audio applications are very popular. Thus, it is important to consider not only single-ended speech quality evaluation but also music quality evaluation. Music and speech are always mixed, such as in Internet radios, where there is usually no clear point when music or speech only will be broadcast. Hence, a general audio quality evaluation is important. Although only speech was considered in this study, this study can be extended to include evaluation of the general audio quality of IP audio applications, including both speech and music signals. Furthermore, highly versatile codecs such as Opus (Valin et al., 2012), which scales from low-bitrate narrowband speech to high-fidelity full-band speech and supports packet loss concealment and the jitter buffer algorithm should be studied.

For VoIP transcription or any other ASR that translates noisy speech transmitted through IP network into text, the ASR model trained on noise-network distorted speech performs

better than the clean-speech-trained model. The ASR model trained on noise-network distorted speech can tolerate a jitter of less than 20% and a packet loss of less than 15% without a decrease in the performance. These results are based on G.722 speech codec without any jitter buffer algorithms and packet loss concealment support. In the next study we will extend this study to include highly versatile codecs like Opus codec (Valin et al., 2012) which can scale from low bitrate narrowband speech to full-band speech, support for packet loss concealment and jitter buffer algorithms.

In this study, the dataset includes 30 sentences, which covers all phonemes in the American English language. However, this dataset is small, which results in a degradation on generalization performance. In future studies, a large dataset with a rich set of utterances and speakers can be considered in order to improve the generalization performance of the ASR model trained on noise-network distorted speech dataset. The training method should also be improved in order to maintain generalization performance while learning noise-network distortion features.

It should be noted that the proposed model does not consider the effect of degrading talking or conversation quality. These include response delay, side-tone, talker-echo or any other two-way interaction features. This study provides an overview of the effect of noise distortions and VoIP-transmission-induced distortions on speech when used as input to ASR. The results of this study can help with network planning for VoIP transcription applications or the deployment of ASR systems, where speech is captured in noisy environments and the transcription is performed remotely.

Bibliography

Alahmadi, M., Pocta, P., & Melvin, H. (2021). An Adaptive Bitrate Switching Algorithm for Speech Applications in Context of WebRTC. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(4). https://doi.org/10.1145/3458751

Alkhawaldeh, R. S., Khawaldeh, S., Pervaiz, U., Alawida, M., & Alkhawaldeh, H. (2019). NIML: Non-intrusive machine learning-based speech quality prediction on VoIP networks. *IET Communications*, *13*(16), 2609–2616. https://doi.org/10.1049/iet-com.2018.5430

Ang, L. Y. L., Koh, Y. K., & Lee, H. P. (2017). The performance of active noise-canceling headphones in different noise environments. *Applied Acoustics*, *122*, 16–22. https://doi.org/https://doi.org/10.1016/j.apacoust.2017.02.005

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 4218–4222.

Barker, J., Marxer, R., Vincent, E., & Watanabe, S. (2015). The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 504–511. https://doi.org/10.1109/ASRU.2015.7404837

Barker, J., Vincent, E., Ma, N., Christensen, H., & Green, P. (2013). The PASCAL CHiME Speech Separation and Recognition Challenge. *Computer Speech and Language*, *27*(3), 621–633. https://doi.org/10.1016/j.csl.2012.10.004

Barker, J., Watanabe, S., Vincent, E., & Trmal, J. (2018). The Fifth "CHiME" Speech Separation and Recognition Challenge: Dataset, Task and Baselines. *Proc. Interspeech 2018*, 1561–1565. https://doi.org/10.21437/Interspeech.2018-1768

Brown, K. L., & George, E. B. (1995). CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*, 105–108. https://doi.org/10.1109/icassp.1995.479284

Catellier, A. A., & Voran, S. D. (2020). Wawenets: A No-Reference Convolutional Waveform-Based Approach to Estimating Narrowband and Wideband Speech Quality. *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.*, *2020-May*, 331–335. https://doi.org/10.1109/ICASSP40776.2020.9054204

Cauchi, B., Siedenburg, K., Santos, J. F., Falk, T. H., Doclo, S., & Goetze, S. (2019). Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network.

IEEE/ACM Trans. Audio Speech Lang. Process., *27*(7), 1151–1163. https://doi.org/10.1109/TASLP.2019.2912123

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., & Johnson, M. (2000). Bllip 1987-89 wsj corpus release 1. In *Linguistic Data Consortium, Philadelphia*.

da Silva, A. P. C., Varela, M., de Souza e Silva, E., Leão, R. M. M., & Rubino, G. (2008). Quality assessment of interactive voice applications. *Computer Networks*, *52*(6), 1179–1192. https://doi.org/10.1016/j.comnet.2008.01.002

Falk, T. H., & Chan, W. Y. (2006a). Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(6), 1935–1947. https://doi.org/10.1109/TASL.2006.883253

Falk, T. H., & Chan, W. Y. (2006b). Enhanced non-intrusive speech quality measurement using degradation models. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*. https://doi.org/10.1109/icassp.2006.1660151

Fu, S.-W., Tsao, Y., Hwang, H.-T., & Wang, H.-M. (2018). Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM. *Interspeech*, 1873–1877.

Furui, S., Maekawa, K., & Hitoshi, I. (2000). A Japanese National Project on Spontaneous Speech Corpus and Processing Technology. *ASR2000-Automatic Speech Recognit. Challenges New Millenium ISCA Tutor. Res. Work.*, *58*(12), 7250–7257. https://doi.org/10.1128/AAC.03728-14

Gamper, H., Reddy, C. K. A., Cutler, R., Tashev, I. J., & Gehrke, J. (2019). Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network. *2019 IEEE Work. Appl. Signal Process. to Audio Acoust.* https://doi.org/10.1109/WASPAA.2019.8937202

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1993). The DARPA TIMIT Acoustic-PhoneticContinuous Speech Corpus CD-ROM. Technical Report NISTIR 4930. *Natl. Inst. Stand. Technol.*

Gysel, P., Pimentel, J., Motamedi, M., & Ghiasi, S. (2018). Ristretto: A Framework for Empirical Study of Resource-Efficient Inference in Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(11), 5784–5789. https://doi.org/10.1109/TNNLS.2018.2808319

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *ArXiv Preprint ArXiv:1412.5567*.

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*(7–8), 588–601.

https://doi.org/10.1016/j.specom.2006.12.006

Hu, Z. G., Yan, H. R., Yan, T., Geng, H. J., & Liu, G. Q. (2020). Evaluating QoE in VoIP networks with QoS mapping and machine learning algorithms. *Neurocomputing*, *386*, 63–83. https://doi.org/10.1016/j.neucom.2019.12.072

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 7304–7308. https://doi.org/10.1109/ICASSP.2013.6639081

Huber, R., & Kollmeier, B. (2006). PEMO-Q-A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(6), 1902–1911. https://doi.org/10.1109/TASL.2006.883259

Hubert, B. (2001). *tc(8) - Linux manual page*. https://man7.org/linux/man-pages/man8/tc.8.html

IEEE. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE No 297-1969*, 1–24. https://doi.org/10.1109/IEEESTD.1969.7405210

INTERNATIONAL TELECOMMUNICATION UNION. (1996). Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*.

INTERNATIONAL TELECOMMUNICATION UNION. (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*.

INTERNATIONAL TELECOMMUNICATION UNION. (2003). Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *ITU-T Recommendation P.835*.

INTERNATIONAL TELECOMMUNICATION UNION. (2004). Single-ended method for objective speech quality assessment in narrow-band telephony applications. *ITU-T Recommendation P.563*.

INTERNATIONAL TELECOMMUNICATION UNION. (2011). Perceptual Objective Listening Quality Assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband. *ITU-T Recommendation P.863*.

Islam, R., Rahman, A., Hasan, N., Hossain, A. N. M. S., Uddin, A. N., & Haque, M. A. (2017). Non-intrusive objective evaluation of speech quality in noisy condition. *Proceedings of 9th International Conference on Electrical and Computer Engineering, ICECE 2016*, 586–589. https://doi.org/10.1109/ICECE.2016.7853988

ITU-T. (2005). G.722.1,“Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss.” *International Telecommunication Union, Geneva, Switzerland*.

ITU-T. (2015). The E-model: a computational model for use in transmission planning. In *International Telecommunication Union*. Recommendation ITU-T G.107. http://handle.itu.int/11.1002/1000/12505

ITU. (2018). *P.863 : Perceptual objective listening quality prediction*. ITU-T Recommendation. https://www.itu.int/rec/T-REC-P.863

Jassim, W. A., & Zilany, M. S. (2019). NSQM: A non-intrusive assessment of speech quality using normalized energies of the neurogram. *Computer Speech and Language*, *58*, 260–279. https://doi.org/10.1016/j.csl.2019.04.005

Jelassi, S., & Rubino, G. (2018). A perception-oriented Markov model of loss incidents observed over VoIP networks. *Computer Communications*, *128*, 80–94. https://doi.org/https://doi.org/10.1016/j.comcom.2018.06.009

Kermanshahi, M. A., Akbari, A., & Nasersharif, B. (2021). Transfer Learning for End-to-End ASR to Deal with Low-Resource Problem in Persian Language. *26th International Computer Conference, Computer Society of Iran, CSICC 2021*. https://doi.org/10.1109/CSICC52343.2021.9420540

Köhn, A., Stegen, F., & Baumann, T. (2016). Mining the Spoken Wikipedia for Speech Data and Beyond. *Proc. Tenth Int. Conf. Lang. Resour. Eval.*

Koizumi, Y., Niwa, K., Hioka, Y., Kobayashi, K., & Haneda, Y. (2017). DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 81–85. https://doi.org/10.1109/ICASSP.2017.7952122

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do Better ImageNet Models Transfer Better? *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, *2019-June*, 2656–2666. https://doi.org/10.1109/CVPR.2019.00277

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., & Shikano, K. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, *9*(4), 357–363. https://doi.org/https://doi.org/10.1016/0167-6393(90)90011-W

Li, B., & Sim, K. C. (2014). A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(8), 1296–1305. https://doi.org/10.1109/TASLP.2014.2329237

Linux Foundation. (2021). *networking:netem [Wiki]*.

https://wiki.linuxfoundation.org/networking/netem

Liu, F., Zhao, X., Zhu, Z., Zhai, Z., & Liu, Y. (2023). Dual-microphone active noise cancellation paved with Doppler assimilation for TADS. *Mechanical Systems and Signal Processing*, *184*, 109727. https://doi.org/https://doi.org/10.1016/j.ymssp.2022.109727

Loizou, P. C. (2013). *Speech enhancement: Theory and practice*. CRC press.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457.

Manocha, P., Xu, B., & Kumar, A. (2021). NORESQA: A Framework for Speech Quality Assessment using Non-Matching References. *Advances in Neural Information Processing Systems*, *27*(NeurIPS), 22363–22378.

Mittag, G., Cutler, R., Hosseinkashi, Y., Revow, M., Srinivasan, S., Chande, N., & Aichner, R. (2020). *DNN No-Reference PSTN Speech Quality Prediction*. https://doi.org/10.21437/Interspeech.2020-2760

Mittag, G., & Möller, S. (2019). Non-intrusive Speech Quality Assessment for Super-wideband Speech Communication Networks. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, *2019-May*, 7125–7129. https://doi.org/10.1109/ICASSP.2019.8683770

Mittag, G., Naderi, B., Chehadi, A., & Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *ArXiv Preprint ArXiv:2104.09494*.

Molesworth, B. R. C., Burgess, M., & Kwon, D. (2013). The use of noise cancelling headphones to improve concurrent task performance in a noisy environment. *Applied Acoustics*, *74*(1), 110–115. https://doi.org/https://doi.org/10.1016/j.apacoust.2012.06.015

Moon, T., Choi, H., Lee, H., & Song, I. (2016). RNNDROP: A novel dropout for RNNS in ASR. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 65–70. https://doi.org/10.1109/ASRU.2015.7404775

Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *INTERSPEECH*.

Mozilla. (2020). *DeepSpeech 0.9.3*. https://github.com/mozilla/DeepSpeech

Ooster, J., Huber, R., & Meyer, B. T. (2018). Prediction of perceived speech quality using deep machine listening. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018-Septe*(September), 976–980. https://doi.org/10.21437/Interspeech.2018-1374

Ooster, J., & Meyer, B. T. (2019). Improving Deep Models of Speech Quality Prediction through Voice Activity Detection and Entropy-based Measures. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2019-May*, 636–640. https://doi.org/10.1109/ICASSP.2019.8682754

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2015-Augus*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Pascual, S., Bonafonte, A., & Serrà, J. (2017). *SEGAN: Speech Enhancement Generative Adversarial Network*. http://arxiv.org/abs/1703.09452

Povey, D., Boulianne, G., Burget, L., Motlicek, P., & Schwarz, P. (2011). The Kaldi Speech Recognition. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Purin, M., Sootla, S., Sponza, M., Saabas, A., & Cutler, R. (2022). Aecmos: a Speech Quality Assessment Metric for Echo Impairment. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2022-May*, 901–905. https://doi.org/10.1109/ICASSP43922.2022.9747836

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., & Omologo, M. (2015). The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 275–282. https://doi.org/10.1109/ASRU.2015.7404805

Rodriguez, D. Z., Rosa, R. L., Almeida, F. L., Mittag, G., & Moller, S. (2019a). Speech Quality Assessment in Wireless Communications with MIMO Systems Using a Parametric Model. *IEEE Access*, *7*, 35719–35730. https://doi.org/10.1109/ACCESS.2019.2902798

Rodriguez, D. Z., Rosa, R. L., Almeida, F. L., Mittag, G., & Moller, S. (2019b). Speech Quality Assessment in Wireless Communications with MIMO Systems Using a Parametric Model. *IEEE Access*, *7*, 35719–35730. https://doi.org/10.1109/ACCESS.2019.2902798

Rousseau, A., Deléglise, P., & Estève, Y. (2012). TED-LIUM: an Automatic Speech Recognition dedicated corpus. *Proc. Eight Int. Conf. Lang. Resour. Eval.*, 125–129. http://www.ted.com

Sharma, D., Wang, Y., Naylor, P. A., & Brookes, M. (2016). A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*, *80*, 84–94. https://doi.org/10.1016/j.specom.2016.03.005

Shi, L., Bao, F., Wang, Y., & Gao, G. (2019). Research on Transfer Learning for Khalkha

Mongolian Speech Recognition Based on TDNN. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 85–89. https://doi.org/10.1109/IALP.2018.8629237

Soni, M. H., & Patil, H. A. (2016a). Novel deep autoencoder features for non-intrusive speech quality assessment. *European Signal Processing Conference*, 2315–2319. https://doi.org/10.1109/EUSIPCO.2016.7760662

Soni, M. H., & Patil, H. A. (2016b). Novel deep autoencoder features for non-intrusive speech quality assessment. *European Signal Processing Conference*, *2016-Novem*, 2315–2319. https://doi.org/10.1109/EUSIPCO.2016.7760662

Sun, L., & Ifeachor, E. C. (2006). Voice quality prediction models and their application in VoIP networks. *IEEE Transactions on Multimedia*, *8*(4), 809–820. https://doi.org/10.1109/TMM.2006.876279

Sun, L., & Ifeachor, E. C. (2002a). Perceived speech quality prediction for voice over IP-based networks. *IEEE International Conference on Communications*, *4*, 2568–2572. https://doi.org/10.1109/icc.2002.997307

Sun, L., & Ifeachor, E. C. (2002b). Perceived speech quality prediction for voice over IP-based networks. *IEEE Int. Conf. Commun.*, *4*, 2568–2572. https://doi.org/10.1109/icc.2002.997307

The FFmpeg developers. (2020). *FFmpeg Documentation*. https://www.ffmpeg.org/ffmpeg.html

Thilo, T., William, T. C., Roland, B., Christian, S., Sporer, T., Beerends, J. G., Catherine, C., Michael, K., Gerhard, S., Karlheinz, B., & Bernhard, F. (2000). PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, *48*(1/2), 3–29.

Uhl, T. (2018). QoS by VoIP under Use Different Audio Codecs. *Proceedings of 2018 Joint Conference - Acoustics, Acoustics 2018*, 311–314. https://doi.org/10.1109/ACOUSTICS.2018.8502317

Valentini-Botinhao, C. (2017). Reverberant speech database for training speech dereverberation algorithms and TTS models, 2016 [dataset]. In *University of Edinburgh*. University of Edinburgh. https://doi.org/https://doi.org/10.7488/ds/2117

Valin, J.-M., Vos, K., & Terriberry, T. (2012). Definition of the Opus audio codec. *IETF RFC6716*, *2*. http://www.hjp.at/doc/rfc/rfc6716.html

Vieira, S. T., Rosa, R. L., & Rodrguez, D. Z. (2020). A speech quality classifier based on tree-cnn algorithm that considers network degradations. *Journal of Communications Software and Systems*, *16*(2), 180–187. https://doi.org/10.24138/jcomss.v16i2.1032

Voiers, W. (1980). Interdependencies among measures of speech intelligility and speech

"Quality." *ICASSP '80. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *5*, 703–705. https://doi.org/10.1109/ICASSP.1980.1170874

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese Broadcast News Corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *10*(2), 219–236.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). *Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR*. https://hal.inria.fr/hal-01163493

Wiener, N. (1949). The Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications. In *Extrapolation, Interpolation, Smoothing Station. Time Ser. with Eng. Appl.* MIT press.

Wuttidittachotti, P., & Daengsi, T. (2017). Subjective MOS model and simplified E-model enhancement for Skype associated with packet loss effects: a case using conversation-like tests with Thai users. *Multimedia Tools and Applications*, *76*(15), 16163–16187.

Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2014). An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE SIGNAL PROCESSING LETTERS*, *21*(1). https://doi.org/10.1109/LSP.2013.2291240

Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., & Nakatani, T. (2015). The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 436–443. https://doi.org/10.1109/ASRU.2015.7404828

Yu, M., Zhang, C., Xu, Y., Zhang, S., & Yu, D. (2021). MetricNet: Towards improved modeling for non-intrusive speech quality assessment. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *4*, 2838–2842. https://doi.org/10.21437/Interspeech.2021-659

Publications

Kumalija, Elhard James, and Yukikazu Nakamoto. 2023. "*MiniatureVQNet: A Light-Weight Deep Neural Network for Non-Intrusive Evaluation of VoIP Speech Quality*" Applied Sciences 13, no. 4: 2455. https://doi.org/10.3390/app13042455

Kumalija, Elhard James and Yukikazu Nakamoto. 2022. "*Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech*" Frontiers in Signal Processing, volume 2, https://doi.org/10.3389/frsip.2022.999457

Kumalija, Elhard James, and Yukikazu Nakamoto. 2020. "*Live Monitoring of Speech Quality of Public Addressing Network Speakers: A Preliminary Study*," In Proceedings of the 3rd ACM Artificial Intelligence and Cloud Computing Conference, pp. 97-101. https://doi.org/10.1145/3442536.3442551