# Voice Quality Experience Evaluation: MOS Laboratory Test and Spectrum Analysis of Failures in VoLTE Calls

Abdel Fadyl Chabi, Matheus Fontinele de Aguiar, Jordan Kalliure S. Carvalho
Vivianne de Aquino Rodrigues, João Vitor da S. Campos, Bruna Maira da S. Fonseca, Janislley O. De Sousa
*Sidia Institute of Science and Technology*
Manaus-AM, Brazil
{abdel.chabi, matheus.fontinele, jordan.carvalho, vivianne.aquino, joao.vitor, bruna.fonseca, janislley.sousa}@sidia.com

*Abstract*—**In recent years, Voice over IP (VoIP) call feature has become increasingly accessible to customers due to advancements in internet services and the emergence of voice call applications. However, identifying the variables that impact Voice over Long-Term Evolution (VoLTE) performance, particularly in quantifying end-user experience in the field and the effects of radio conditions and IP impairments on voice quality as measured by the Mean Opinion Score (MOS), presents challenges for carriers. MOS is a widely used metric for evaluating voice quality, and there is a significant commitment from both mobile device manufacturers and carriers to ensure superior voice quality during voice calls. To this end, MOS experiments are performed to evaluate the reliability of VoLTE calls, which is currently the best approach for measuring voice quality. In this study, we present MOS experimentation results in laboratory environments to homologate 146 different smartphone models. As results, we highlight the challenges associated with MOS testing in VoLTE calls under controlled conditions and discuss the primary issues found and how they were addressed. These experimental analyses offer substantial opportunities for enhancing the design and operation of audio quality during VoLTE calls and detail potentially improvements for 5G VoNR calls.**

***Keywords — MOS; POLQA; VoLTE; AMR; Mobile Network Tests***

## I. Introduction

The availability of VoIP has increased in recent years due to advancements in internet services and the development of voice call apps [1]. To ensure software quality on mobile devices, carriers must implement testing phases and structured processes. As suggested by the International Software Testing Qualifications Board (ISTQB) [2], thorough testing and documentation can reduce the risk of failure during operation. Additionally, MOS test is crucial in determining the effectiveness of the testing phase, as it directly relates to the user experience [3]. It is the responsibility of smartphone manufacturers and carriers to ensure high-quality VoLTE services for users [4].

The human voice is naturally an analog signal [5], not defined by physical or mathematical equations. Similarly, the voice quality is a subjective measure. For many years, in a telephone network, this measure was also determined subjectively, based on evaluations by users who assigned scores from 1 to 5, representing their perceptions of the quality of a given audio segment [6]. To determine the quality of the voice, the scores provided by the test participants are subjected to an arithmetic mean calculation, called MOS [3].

In the research conducted by Sung *et al.* [7], a study of voice quality was conducted using eight different scenarios in order to measure the quality of VoLTE services using Perceptual Objective Listening Quality Assessment (POLQA) models under varying signal strengths of the reference signal. The results of the study indicate that audio quality is affected on certain brands of mobile devices and comparisons were made between the POLQA-NB and POLQA-WB algorithms. Additionally, in the study [8], an investigation of various technologies was conducted to provide high-quality cellular voice onboard trains. To determine the voice quality, a setup for measuring the Perceptual Evaluation of Speech Quality (PESQ) was developed.

The quality of VoLTE calls is commonly assessed using MOS measured by PESQ or POLQA algorithms [9]. MOS scores that fall below the acceptable range set by the carrier indicate voice quality issues, which can arise from various factors, including poor network or channel conditions, audio processing in the Audio Digital Signal Processor (ADSP), and packet loss or misses in the system [7]. When performing MOS experiments in a laboratory environment, additional factors that can lead to MOS degradation must be considered, such as device hardware limitations and codec gain conditions. The importance of researching MOS experimentation and optimization in VoLTE calls lies in the need to ensure excellent voice quality, which is critical for the success of the technology and customer satisfaction.

The objective of this study is to advance the field of Audio quality testing using MOS for VoLTE calls. An experience report is presented on MOS test results obtained in a telecommunications laboratory, called ATLAS, which is dedicated to the homologation process for mobile devices in Latin America. The focus of this experience report is to share the gains and strengths achieved in fixing MOS failures and the results are presented on the management and analysis of failures in a controlled environment during VoLTE call tests. The spectrum analysis was performed using specialized equipment to measure the strength and characteristics of the radio frequency signals in the environment. These results can be used to identify potential sources of interference and to evaluate the impact of this interference on the MOS. This work may be of a great interest to Mobile Service Providers and will contribute to adjust and calibrate their operations in order to certify the best quality of voice calls over VoLTE.

The paper is organized as follows: In Section II, the background on LTE networks, VoLTE calls and the POLQA standard for evaluating MOS experiments are presented. The testing methodology used in the laboratory is outlined in Section III. Section IV describes the experimental methodology and environment employed in this study. The data collected in laboratory experiments and the challenges encountered in addressing them are presented in Section V. In Section VI, the statistics and resolutions related to MOS failures are discussed. Finally, the conclusion is presented.

## II. Background

Telephone calls are traditional services offered by telephony carriers' networks in Global System for Mobile Communications (GSM), Wideband Code Division Multiple Access (WCDMA), and Long-Term Evolution (LTE) technologies. The network architectures of GSM and WCDMA are based on Circuit Switching (CS), in contrast to LTE, which has a Packet Switching (PS) based architecture [10]. According to [11], the key difference between CS and PS mode is: In CS mode, there is a physical channel reserved until data transmission starts. It means that a dedicated communication path is established between the sender and the receiver. When CS is compared with PS, is eliminated the need to transmit the message in a dedicated path. In PS, the data is split in packets and grouped together. Each packet is routed from the source to the destination individually.

Since the beginning of GSM technology, voice telephone calls occur in the CS domain. Therefore, since LTE technology has a PS-based architecture, the IP Multimedia Subsystem (IMS) was introduced as a resource of the LTE network to enable the transmission of voice calls using the IP protocol [12], as shown in Fig. 1. The VoLTE has the characteristic of allowing the transmission of high-definition voice and faster call setup time compared to voice calls in GSM and WCDMA networks [13]. In this scenario, the MOS test evaluates the voice quality in a call, and thus, voice calls for laboratory trials are made in VoLTE.

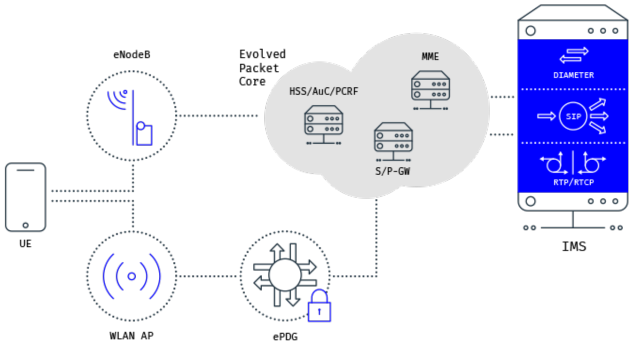

Figure 1. High-level LTE Architecture and IMS core [14]

### A. Voice Quality Standard

POLQA is the current voice quality testing standard for mobile devices, established by the International Telecommunication Union (ITU) in ITU-T P.863 release of 2010 [9] and is an evolution of the PESQ in the ITU-T P.862 release of 2000. POLQA was specifically developed for the requirements of HD voice super wideband, 3G, VoLTE, and VoHSPA. The POLQA algorithm compares a reference audio signal (input signal) with a distorted output signal that passed through a communication system, where encoding, and decoding processes took place, and considering Radio Frequency (RF) losses [15]. The output of the algorithm is a prediction of the perceived quality of the output signal as it would be similarly heard by a group of people in a subjective listening test.

Fig. 2 illustrates the POLQA workflow during the MOS test. The operations that modify the characteristics of the reference signal and the distorted signal are used for the idealization process. This subjective test is performed using listening or conversation method and the results are often reported on a scale of 1 to 5, with 1 being poor and 5 being exceptional.

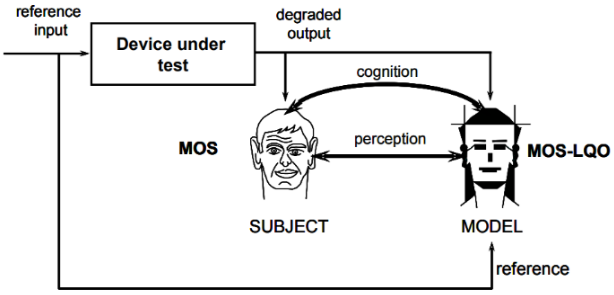

Figure 2. POLQA workflow during the MOS test [9]

The cognitive model computes six quality indicators: Frequency Response Indicator (FREQ), Noise Indicator (NOISE), Room Reverb Indicator (REVERB) and three indicators that describe the internal difference in the Time-Pitch-Loudness domain. These indicators provide a comprehensive evaluation of the speech quality based on various factors such as frequency response, presence of noise and reverb, and differences in the time-pitch-loudness domain. They enable a detailed assessment of speech quality and can be used to identify areas for improvement in the communication system.

These indicators are combined to provide an objective listening quality using Mean Opinion Score - Listening Quality Objective (MOS-LQO). It calculates the score by objectively predicting the quality of a listening-only test scenario. Also, POLQA always expects a clean reference signal (noise-free).

### B. Speech Coding Techniques: AMR-NB and AMR-WB

Speech coding is essentially the process of converting an initial audio signal into a more compressed form so that it can be transmitted using fewer bits, while still using techniques so that the original signal can be recovered without damage [16]. Furthermore, two parameters help to classify speech coders: the bit rate at which the coders' output has a reasonable quality output, and the coding technique used.

The Adaptive Multi-Rate Codec (AMR) is a voice coding algorithm originally developed for GSM systems and is now adopted as the standard voice codec by 3GPP. AMR is a combination of voice codecs and channels activated and controlled through signaling to provide better speech quality under transmission errors and noise [17]. Most smartphones today use the AMR audio format for audio storage. There are two types of AMR codecs. The first is Adaptive Multi-Rate Narrowband (AMR-NB), which is a narrowband voice codec that can operate at various bit rates, ranging from 4.75 kbit/s to 12.2 kbit/s [17]. The second is Adaptive Multi-Rate Wideband (AMR-WB), which is a wideband voice codec that operates at bit rates ranging from 6.6 kbit/s to 23.85 kbit/s [17].

Most voice coders in mobile communications use the AMR-NB technique [18]. However, with the evolution of mobile networks, AMR-WB technique has become highly desirable, as the most current mobile network technologies use ever-increasing frequencies and bandwidths. Table I refers to the methods used to compare the performance of different AMR codecs. Working with greater bandwidths delivers high intelligibility, naturalness and quality speech. To ensure audio quality meets the standards set forth by carriers, it is crucial to employ both techniques in MOS laboratory tests. For laboratory tests, the MOS measure must exceed 3.5 in accordance with the carrier requirement.

Table I
AMR COMPARATIVE ALGORITHMS

| Codec | Audio Bandwith | Technology | Bitrates (Kbps) |
|---|---|---|---|
| *Enhanced Full-Rate (EFR)* | Narrow Band | 2G | 12.2 |
| *AMR Narrowband (AMR-NB)* | Narrow Band | 3G/4G | 4.72 - 12.2 |
| *AMR Wideband (AMR-WB)* | Wide Band | HD Voice in 3G/4G | 6.6 - 23.85 |
| *Extended AMR-WB (AMR-WB+)* | Full Band | Full HD Voice in 3G/4G | 6 - 48 |
| *Enhanced Voice Service (EVS)* | Full Band<br>Super Wide Band<br>Wide Band<br>Narrow Band | Super HD Voice in 4G | 5.9 - 128 |

## III. MOS Testing Methodology

### A. MOS Laboratory

The Advanced Telecommunication Laboratory in Amazon Sidia (ATLAS) plays a crucial role in assisting the software development team by conducting telecommunications tests on Android mobile devices as per the established protocols of various carriers. These tests are conducted at different stages of the PLC, defining the various project stages. Fig. 3 depicts the key milestones in software development pertaining to the MOS testing stage. This analysis is conducted from both the perspective of mobile device manufacturers, also know as Original Equipment Manufacturer (OEM), and carriers.

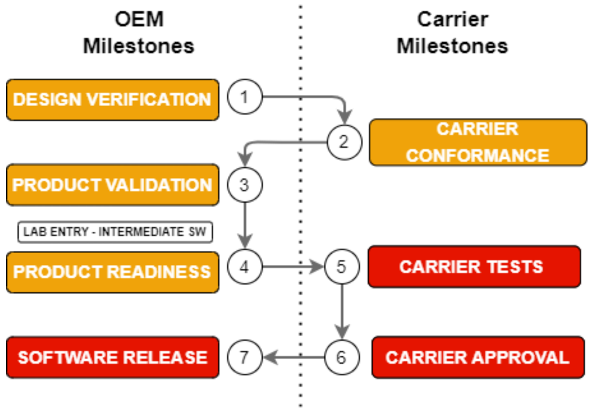

Figure 3. PLC Milestones used for MOS tests in ATLAS, including OEM and Carrier ones

The diagram on Fig. 3 depicts the processes and milestones involved in the release of new software to consumers. It starts with the Design Verification stage, which ensures that the designed software product meets the established design specifications. Once approval is obtained, the product specification information is shared with the carrier to analyze and ensure the device's safe operation on it. The carrier then grants the OEM permission to proceed with product development in order to fulfill the requirements, a process known as Carrier Conformance. Subsequently, the Product Validation stage commences, where validation activities for the developed project are carried out to identify and rectify any possible software (SW) or hardware (HW) failures. At the conclusion of this stage, an intermediate software version is released, known as the lab entry milestone. This is followed by the Product Readiness stage, where the ATLAS team and other test teams commence their work. In this final stage, tests are focused on the carrier's specific features until the final version of the product is reached. At this point, the Carrier Tests stage begins, where the OEM provides the final software for the carrier to carry out their tests. Upon finalizing and approving all tests, the carrier approves the HW and SW of the OEM (Carrier Approval). Finally, after getting carrier permission, the OEM releases the new software to users.

Carriers aim to optimize their network resources to enhance user satisfaction, which involves controlling audio and video transmission quality. MOS testing is crucial for evaluating the user experience [19], and ATLAS models are validated through communication protocol tests in a controlled environment that simulates different mobile network scenarios, including MOS experiments. MOS failures can impact the mobile device homologation process with carriers, highlighting the importance of promptly addressing any identified failures. To perform accurate analysis, it is crucial to understand the operation and techniques used to avoid inconsistent results.

### B. MOS Test Cases

The MOS test cases utilize the AMR technique to optimize speech encoding. AMR is part of the Linear Predictive Coding (LPC) encoder class, which uses multiple bit rates and an error concealment mechanism to circumvent errors and packet losses during transmission [20]. The MOS is conducted in the ATLAS laboratory for Latin American carriers, using AMR-WB and AMR-NB techniques. AMR-WB integrates a single codec with nine different rates ranging from 23.85 kbit/s to 6.6 kbit/s, while AMR-NB has eight rates ranging from 12.2 kbit/s to 4.75 kbit/s [21], as shown in Table II. With AMR, the encoder is able to change its bitrate every 20 ms corresponding to a speech frame as received commands from the encoder.

Table II
CODEC RATE USED IN MOS TEST

| Codec | Bit Rate (kbps) |
|---|---|
| AMR-WB | 23.85 / 23.05 / 19.85 / 18.25 / 15.85 / 14.25 / 12.65 / 8.850 / 6.600 |
| AMR-NB | 12.20 / 10.20 / 7.950 / 7.400 / 6.700 / 5.900 / 5.150 / 4.750 |

### C. MOS Experimentation Settings

To conduct MOS tests, a mobile network simulator and an audio analyzer infrastructure are required. This is achieved by generating an LTE signal with an enabled IMS server, which is done using the equipment *CMW500 Universal Communication Tester*, allowing the cell phone to register on the network. Then, a VoLTE call is established in the Device Under Test (DUT), and finally, audio measurements are taken using the POLQA algorithm, with the help of the audio analyzer *UPV Audio Analyzer* from Rohde&Schwarz.

*1) Hardware Setup:* The CMW500 is the Network generator and it is connected to the UE's RF antenna via cable; while the UPV, which is the audio analyzer, is connected to the UE's audio input via cable.

*2) Software Setup:* The software used and their minimum versions for MOS tests are described in Table III.

Table III
SOFTWARE CONFIGURATION USED IN MOS TEST [21]

| Equipment | Specification |
|---|---|
| CMW500 | Base Firmware $\geq$ 3.5.131<br>Data Application Unit firmware $\geq$ 3.5.50<br>LTE firmware $\geq$ 3.5.50<br>Audio speech firmware $\geq$ 3.5.30 |
| UPV | UPV Firmware $\geq$ 4.0.4<br>POLQA_CAL_macro $\geq$ 1.2.0 |

### *D. Test Execution Method*

This study verifies audio quality in two different scenarios: downlink (DL) and uplink (UL) audio transmitted through cellular voice calls. Figure 4 illustrates the test flow for both scenarios, with the blue and red lines representing the DL and UL, respectively.

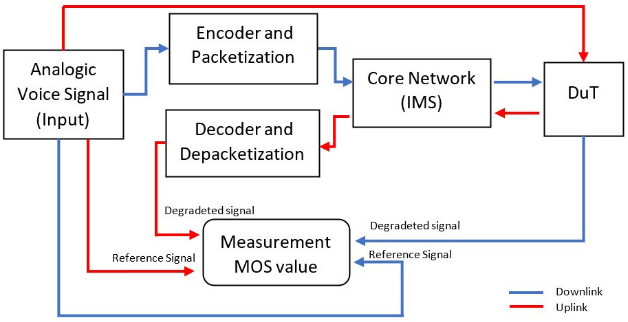

Figure 4. Flow Diagram of MOS tests for Downlink and Uplink

The measurement process begins in the downlink, where the cellular phone is already connected to the LTE network via RF (Radio Frequency) and registered in IMS. In this stage, the DUT receives a call from the virtual client of CMW500, using AMR-WB codec with the test audio, which undergoes the digitization process. The degraded signal is then sent to UPV through the P2 connection, where the MOS calculation is performed by comparing it with the reference signal using the POLQA algorithm. The uplink measurement stage, illustrated in red in Figure 4, involves the DUT receiving the test audio via the P2 connection during an active audio call to the virtual client of CMW500. The audio signal is digitized and transmitted to the IMS server, where it undergoes demodulation and decoding processes, resulting in a degraded signal. This degraded signal is then sent to UPV, where the MOS calculation is performed by comparing it with the reference signal using the POLQA algorithm.

The POLQA method is used to evaluate audio quality, allowing comparisons between the reference signal and the signal reduced by the DUT. The algorithm result is a forecast of the perceived quality of a degraded signal as determined by subjective listening. Furthermore, POLQA measurements are conducted for both downlink and uplink scenarios, covering all codec rates in Table II.

The prediction output of audio quality is evaluated using a MOS score ranging from 1 to 5, corresponding to quality levels ranging from poop to excellent. The evaluation process involves taking 5 POLQA measurements for each codec rate in both scenarios, and the average value is used as the numerical indicator of audio quality. To meet the approval criteria, the numerical value must be at least 3.5.

## IV. CASE STUDY: MOS TEST

The case study presented analyzes a MOS test for AMR-WB Uplink with a failure occurrence. Multiple voice calls were made to the DUT, and measurements were taken during the MOS experiment. The report generated for the 5 measurements is shown in Table IV, as described in section III. The POLQA algorithm's perception for the model under analysis is very close to the limit value established by the carrier's requirement, but not enough to get approval.

Table IV
VALUES OBTAINED IN THE FAILURE OF THE MOS TEST

| Experiment Item | Avg. Delay | MOS |
|---|---|---|
| POLQA Measurement: 001 | 432.0 ms | 3.5517 |
| POLQA Measurement: 002 | 431.9 ms | 3.5531 |
| POLQA Measurement: 003 | 431.9 ms | 3.5073 |
| POLQA Measurement: 004 | 431.8 ms | 3.3873 |
| POLQA Measurement: 005 | 431.8 ms | 3.5000 |

Even though some of the values obtained via POLQA are above the established minimum, the arithmetic mean obtained from the 5 measurements was 3.4998, which is less than 3.5, resulting in failure for the AMR-WB codec rate in UL. In MOS testing, it is common to see voice quality decreasing as codec bitrates are increased. However, for the case presented, the audio quality value is already low from lower frequencies, confirming the UL failure of the model under test.

The study was conducted to investigate the cause of low scores obtained during the MOS test performed. A preliminary analysis revealed a high level of noise in the signal sent to the IMS server by the DUT. Further analysis showed that unoptimized audio parameters, specifically the preprocessing modules and Tx gains, were responsible for the noise. To address this issue, the audio parameters were adjusted until the audio signal reached the range of -12 dB to -18 dB. The experiment was then executed again, and measurements were collected. The new results are presented in Table V. When starting the validation with the new audio configuration applied, it was possible to notice that the measurements performed for each codec were much superior when compared to the measurements obtained in the reported failure. With that, we can verify that the MOS values obtained for the highest codec rates were above 4, confirming that the parameter optimization was sufficient to guarantee the approval of the model regarding the carrier's requirements.

Table V
VALUES OBTAINED IN THE APPROVAL OF THE MOS TEST

| Experiment Item | Avg. Delay | MOS |
|---|---|---|
| POLQA Mensurement: 001 | 463.6 ms | 4.1680 |
| POLQA Mensurement: 002 | 463.6 ms | 4.3756 |
| POLQA Mensurement: 003 | 463.7 ms | 4.3155 |
| POLQA Mensurement: 004 | 463.7 ms | 4.3555 |
| POLQA Mensurement: 005 | 463.7 ms | 4.3169 |

Figure 5 shows the reference audio signal in the time domain, which is the input signal to be compared with the degraded signals during the experiments. This signal is reproduced during the call to allow the equipment to analyze the quality of the signal received by the device during the experiment.

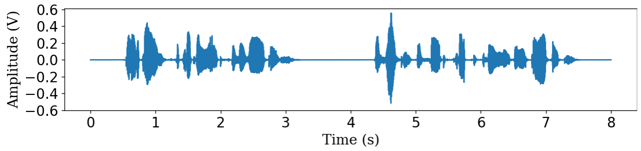

Figure 5. Reference audio signal (input), in the time domain

Initially, the device fails the MOS with a score of 3.4998. In this experiment, the degraded audio signal represented in Figure 6 was sent for MOS measurement, which is also in the time domain. It is possible to analyze in the signal that, at certain times, the audio has a large difference in amplitude when compared to the reference signal in Figure 5.

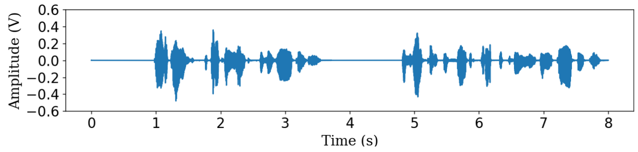

Figure 6. Degraded audio signal that resulted in failure, in time domain

For a better visualization of the failure, Figure 7 shows a graph in the frequency domain comparing the spectrum of the measured signal that failed the experiment with the reference signal. It can be seen that the failure signal is slightly different from the reference signal for frequencies below 1 kHz. In addition, the measured signal suffered a large distortion at frequencies close to 4 kHz. As for frequencies greater than 7 kHz, nothing was captured. After implementing the new audio parameters to the software, the experiment was run again and the behavior of the voice signal sent by the DUT to the IMS server is shown in Figure 8, which is in the time domain and illustrates the new audio signal captured after rerunning the experiment. It is verified that the signal is closer to the reference, where the same codec rate was used, showing that there was a smoothing of the previously captured noise.

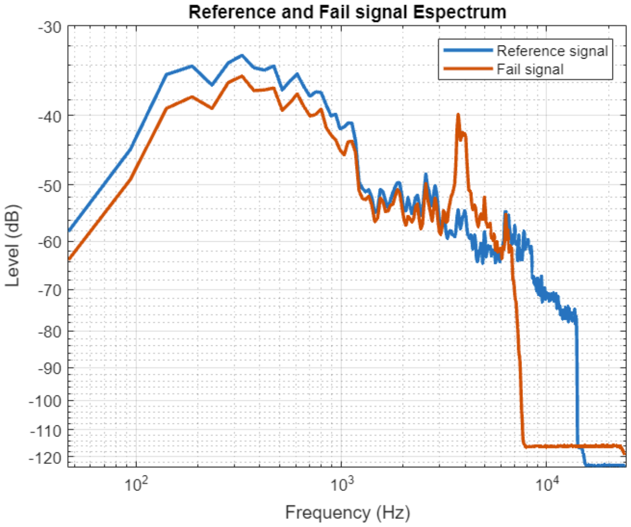

Figure 7. Frequency domain spectrum of the reference audio signal compared to the degraded signal that resulted in the failure

Figure 9 shows the relationship between the reference and pass signals in the frequency domain, which shows the noise reduction ratio for frequencies above 1 KHz and no abnormality in the signal amplitude at frequencies close to 4 kHz.

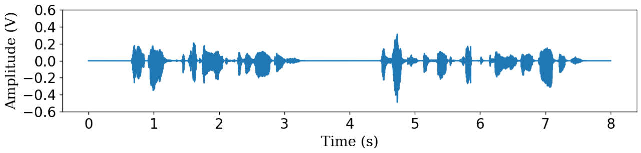

Figure 8. Degraded audio signal (output) that resulted in passing the experiment, in the time domain

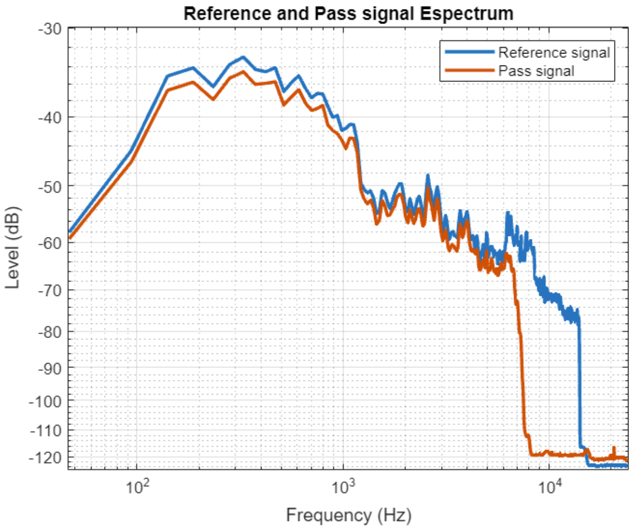

Figure 9. Frequency domain spectrum of the reference audio signal compared to the signal that resulted in the pass

Even though the signals are closer, it is observed that there is still noise in the signal measured by the UPV, but the optimization was enough to guarantee approval in the MOS test according to the POLQA measurements which resulted in an average of 4.3063. As shown in Figure 9, the measured signal was very close to the reference signal for frequencies up to 6 kHz, without suffering major distortions in relation to the reference, which was sufficient to obtain MOS test approval.

## V. MOS Performance Evaluation

Issues with voice quality are frequently reported when the MOS measured by PESQ or POLQA algorithms are low and do not meet the acceptable parameters required by the carrier. The MOS scores can be degraded for several reasons, including network conditions, channel conditions, audio processing, and packet loss/miss. This degradation can occur in UL, DL, or both.

According to laboratory MOS execution tests performed in this study, MOS scores can be degraded due to various factors, including:

- Network conditions: Such as when the experiment is conducted in an area with a weak signal;
- Channel conditions: Such as when the experiment is conducted in an area with poor channel conditions that may cause fading;
- Audio processing: Due to pre and postprocessing modules that are present inside the Audio Digital Signal Processor (ADSP);
- Packet loss/packet miss: This occurs due to packet losses in the network or packet misses in the system.

When conducting MOS tests in a laboratory environment, two additional reasons for MOS score degradation must be considered. These include:

- Performance reduction due to device hardware limitation: In recently released devices, the P2 audio connector output is connected to the USB-C connector used for charging the device's battery, which may cause interference during voice calls;
- Device or Codec gain conditions: During the software development phase of a mobile device, the audio parameter values may fall outside of the standards for transmission (Tx) and/or reception (Rx) gains due to each device's unique HW and SW, resulting in a distinctive behavior in the network.

In case the MOS score is affected by unoptimized audio parameters, developers can take several optimization steps. They can try different preprocessing modules by enabling/disabling them until the one causing the low score is identified through trial and error. Additionally, they can adjust the Rx or Tx gain and volume blocks to keep the signal within the range of -12 dB to -18 dB. Multiple audio configuration files are generated to test and approve the code snippet that fixes the MOS issue. Once the problem is resolved, a new software version is created with the implemented improvements.

## VI. Results and Discussion

In the period from 2020 to 2022, ATLAS lab conducted MOS experiments to homologate 146 Samsung smartphone models and found 12 critical failures. Figure 10 depicts the correlation between the total number of tested models and the number of failures detected.

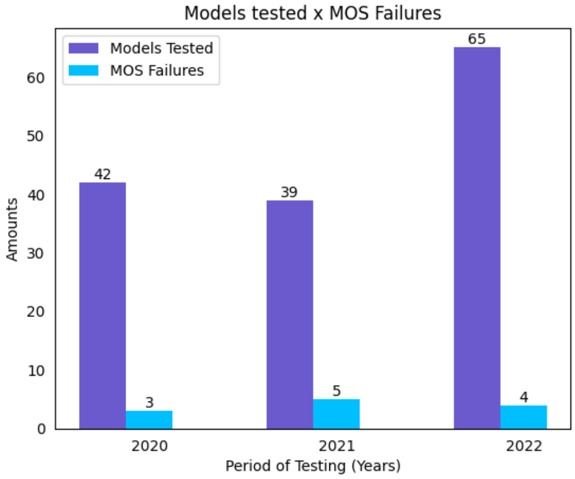

Figure 10. Quantity of models tested and MOS failures found at ATLAS laboratory from 2020 to 2022

Over the years, there has been a growing trend in the demand for MOS tests, which reflects the increasing importance that manufacturers place on ensuring a high standard of quality in voice calls. When failures were reported, they were addressed with a focus on three possible outcomes:

- Correction or optimization of the audio parameters in the software related to voice calls. This may involve adjusting the preprocessing modules and Tx gains, as well as other parameters, to achieve a better MOS score;
- Device hardware failure may be the cause of low MOS scores, such as a performance limitation due to the hardware design of specific models;
- Abnormal behavior of the device may be caused by a chipset incompatibility with the test equipment, which would require the use of different equipment to perform the experiments.

Figure 11 shows the percentages of results found in ATLAS lab experiments. 66% of the failures were resolved by optimizing or correcting audio parameters related to voice calls, indicating that this is the most common solution for MOS failures. This is because each device has its own components and signal capture level, making audio quality improvement the primary focus of analysis.

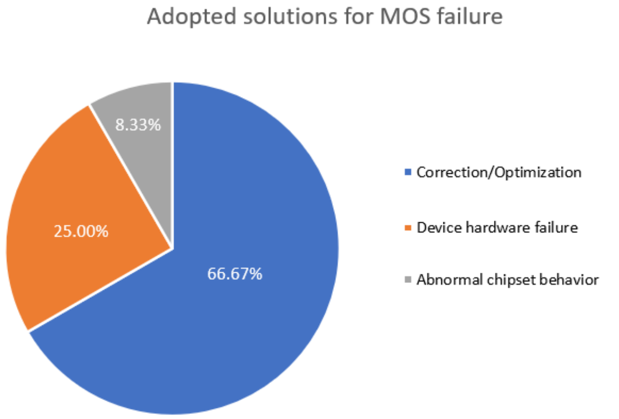

Figure 11. Percentage of results for MOS failures found at the ATLAS lab

The results with hardware failure account for 25% of the total number of failures, which is significant since ATLAS works closely with carriers for validation. In this case, the carrier is informed about the behavior of the device during the conducted experiments, where the devices are connected to the test equipment that analyzes the quality of voice calls. The devices that showed this result suffered interference when using the USB-C headphone jack for the conducted connection. In some cases, especially when validating devices with new chipsets on the market, tests may result in unexpected behavior by the test equipment, leading to false failures, which occurred in 8% of the failures found.

For carriers, audio quality assurance is critical for marketing since these procedures were implemented over an IP/IMS network to assure the greatest user experience for HD Voice conversations, and the quality will vary based on the input audio level. Low or high audio levels can have an influence on non-linear portions of the end-to-end system, thus determining the audio level that offers the optimum speech quality and POLQA score is critical.

## VII. Conclusion

The improvement of mobile network technologies has provided better resources for high-quality services for users. VoLTE voice call services undergo critical tests for audio quality, ensuring good performance in measurements using the POLQA standard in the MOS experiments. However, voice quality measurements using the POLQA algorithm are sensitive to the audio level transmitted and the environment, directly affecting the accuracy of the measurements obtained. MOS scores can be degraded due to various factors such as network conditions, channel conditions, audio processing, and packet loss/miss. This study demonstrate an experimental study with observations performing MOS testing for mobile voice services.

The main challenge faced in conducting MOS tests, as identified in this study, is the high sensitivity of the POLQA algorithm to the audio level and environment, which directly impacts the accuracy of the measurements obtained. It is crucial to determine the optimal audio level to obtain the best

POLQA score, enabling device engineers to ensure the best speech quality for end-users across the network. In Section IV, the case study compares the signals measured during experiments to the reference signal, revealing that the failed signal suffered significant distortion between 3 kHz and 7 kHz, leading to a low MOS score and test failure. On the other hand, the approved signal remained close to the reference signal across the frequency spectrum. Section V emphasizes that MOS scores can be degraded by various factors, such as network conditions, channel conditions, audio processing, and packet loss/miss, and all of these factors need to be addressed to ensure acceptable audio quality during VoLTE calls.

As for dealing with the MOS failures found in the ATLAS laboratory from 2020 to 2022, 66% of the total failures were resolved by correcting and optimizing audio parameters related to voice calls. The remaining 34% of failures were due to hardware limitations and chipset-related issues. The need to adjust audio parameters is quite recurrent since each model has hardware with its components, which can influence the MOS test results. To ensure high-quality voice services are delivered to users, MOS tests should be conducted as early as possible in relation to the device's approval by carriers. Through the results of this study, we present significant insights for improving the design and operation of audio quality during VoLTE calls, as well as details prospective enhancements for 5G VoNR calls.

## Acknowledgment

The authors are grateful for the support offered by SIDIA R&D Institute in the Model 5G project. This work was partially supported by Samsung, using resources of Informatics Law for Western Amazon (Federal Law No. 8.387/1991). Therefore, this work disclosure is in accordance with as foreseen in article No. 39 of number decree 10.521/2020.

## References

[1] S. A. Lonkar and K. Reddy, "Analysis of audio and video quality of voice over lte (volte) call," *International Journal of Information Technology*, pp. 1–14, 2020.

[2] ISTQB, *Certified Tester Foundation Level (CTFL) Syllabus*, 2021. [Online]. Available: https://istqb-main-web-prod.s3.amazonaws.com/media/documents/ISTQB-CTFL_Syllabus_2018_v3.1.1.pdf

[3] C. J. d. SANTOS, "Seleção de codificadores de voz (codecs) através do balanceamento entre qos e qoe auxiliado por algoritmo baseado em análise hierárquica de processos (ahp)," Master's thesis, Universidade Federal de Itajubá, Itajubá, 2019.

[4] *P.863.1: Application guide for Recommendation ITU-T P.863*, International Telecommunication Union (ITU) Std. ITU-T Rec. P.863.1, 2019. [Online]. Available: https://www.itu.int/itu-t/recommendations/rec.aspx?rec=13966

[5] M. Lacuška and T. Peráček, *Trends in Global Telecommunication Fraud and Its Impact on Business*. Cham: Springer International Publishing, 2021, pp. 459–485. [Online]. Available: https://doi.org/10.1007/978-3-030-62151-3_12

[6] R. L. Frigotto *et al.*, "Análise da qualidade de áudio no serviço móvel aeronáutico," Master's thesis, Universidade Tecnológica Federal do Paraná, 2018.

[7] Y. C. Sung, Y.-S. Ho, Y.-B. Lin, J.-C. Chen, and H. C.-H. Rao, "Voice/video quality measurement for lte services," *IEEE Wireless Communications*, vol. 25, no. 4, pp. 96–103, 2018.

[8] H. M. Ylander, "Alternative technologies for providing voice services onboard trains," Master's thesis, Chalmers University of Technology, 2019.

[9] *P.863: Perceptual objective listening quality prediction*, International Telecommunication Union (ITU) Std. ITU-T Rec. P.863, 2018. [Online]. Available: https://www.itu.int/itu-t/recommendations/rec.aspx?rec=13570

[10] A. Elnashar and M. A. El-Saidny, *Practical Guide to LTE-A, VoLTE and IoT: Paving the way towards 5G*. John Wiley & Sons, 2018.

[11] S.-S. Manfred, "Circuit switching versus packet switching," *International Journal of Open Information Technologies*, vol. 3, no. 4, pp. 27–37, 2015.

[12] A. Elnashar and M. A. El-saidny, *Introduction to the IP Multimedia Subsystem (IMS)*. Wiley Telecom, 2018, pp. 87–157.

[13] B. Krasniqi, G. Bytyqi, and D. Statovci, "Volte performance analysis and evaluation in real networks," in *2nd Internetional Balkan Conference on Communications and Networking*, 2018.

[14] Mobileum. Ims and volte testing. [Online]. Available: https://www.mobileum.com/products/testing-and-monitoring/domestic-network-testing/ims-and-volte-testing/

[15] ANACOM, *Mobile Communication Systems*, 2017. [Online]. Available: https://www.anacom.pt/streaming/MetodologiaFinal_QoS_234G_deci sao16junho2017.pdf?contentId1̄412583&fieldĀTTACHED_FILE

[16] A. Gupta and S. Nigam, "A review on different types of lossless data compression techniques," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2021.

[17] D. Choudhary and A. Kumar, "Study and performance of amr codecs for gsm," *IJARCCE*, pp. 8105–8110, 10 2014.

[18] K. J. Byun, I. S. Eo, H. B. Jeong, and M. Hahn, "Real-time implementation of amr and amr-wb using the fixed-point dsp for wcdma systems," in *2006 IEEE International Symposium on Consumer Electronics*, 2006, pp. 1–6.

[19] C. B. MOREIRA, "Avaliação de qualidade de experiência e consumo de energia em transmissão adaptativa de vídeos em dispositivos móveis," Master's thesis, Universidade Federal de Pernambuco, 2016.

[20] Z. Qiao, L. Sun, N. Heilemann, and E. Ifeachor, "A new method for voip quality of service control use combined adaptive sender rate and priority marking," in *2004 IEEE International Conference on Communications (IEEE Cat. No. 04CH37577)*, vol. 3. IEEE, 2004, pp. 1473–1477.

[21] R. . Schwarz, "Voice over lte (volte) speech quality measurements," pp. 1–50. [Online]. Available: https://scdn.rohde-schwarz.com/ur/pws/dl_downloads/dl_application/application_notes/1ma204/1MA204_9e.pdf