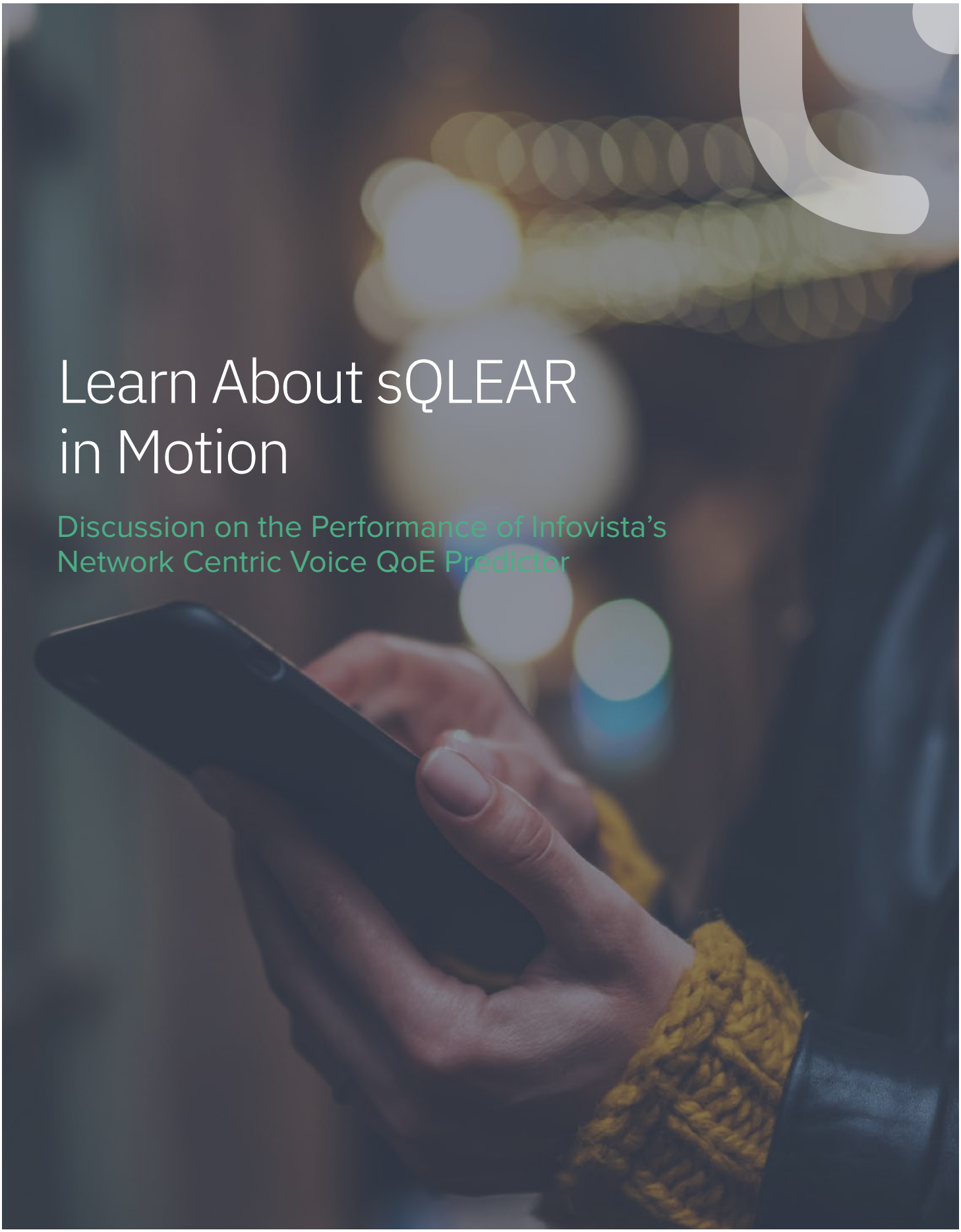
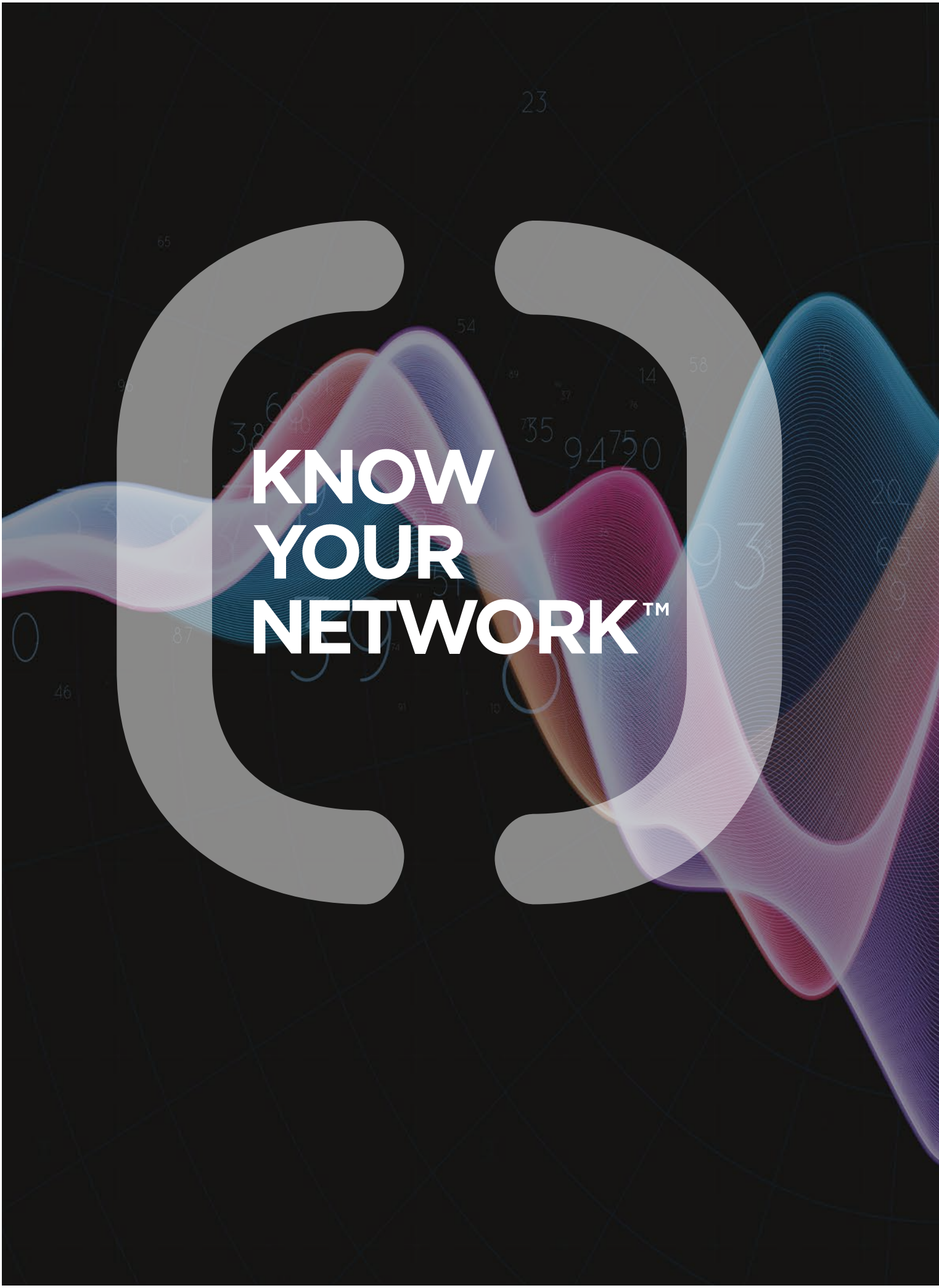
infovista

KNOW YOUR NETWORK™



Learn About sQLEAR in Motion

Discussion on the Performance of Infovista's Network Centric Voice QoE Predictor


KNOW
YOUR
NETWORK™

Content

How does sQLEAR work in a live networks test set-up

Overview

Maintaining the voice service Quality of Experience (QoE), reducing subscriber churn to OTT providers, and growing voice revenue through VoLTE expansion, while optimizing CAPEX/OPEX are key concerns for mobile network operators. Therefore, obtaining accurate, controlled, and easy to implement voice QoE predictors, as well as securing the ability to act in nearly real-time on network centric issues are crucial for enabling cost efficient, optimized network operations that will meet and maintain customer expectations and demands. InfoVista's sQLEAR voice QoE predictor is specifically designed to answer these concerns and goals, benefiting MNOs and regulators alike.

Infovista's sQLEAR is the first industry solution to utilize a combination of machine learning, network/client/codec information and reference audio for the assessment of transmission network impact on speech quality for mobile packet-switched voice services. Therefore, sQLEAR is the first intrusive parametric voice QoE predictor in the industry, which can be used for high definition (super wide and full band) voice QoE testing across 4G/LTE and 5G networks, both carrier (a.k.a VoLTE, VoNR) and OTT (e.g. WhatsApp). This white paper describes sQLEAR concept and its benefits, sQLEAR operational mode and test set up. In addition, the paper discusses sQLEAR performance in lab and field trials, with focus on carrier voice service (a.k.a VoLTE EVS-CA/AMR-IO and AMR-WB) use cases.

Concept

Infovista machine learning based QoE predictor is developed following the new ratified ITU-T P.565 framework's procedures and requirements. sQLEAR uses ITU-T P.565 provided reference speech sample, learning and validation test files, as well as additional learning and validation data sets collected in real live VoLTE and voice generic OTT (WhatsApp based) network scenarios. During the development (a.k.a learning/training and validation) phase (Figure 1) the following information is used to create the inputs (a.k.a features) of sQLEAR ML algorithm: codec (e.g. bit rate, bandwidth, and special modes of EVS codec - Channel Aware CA and AMR-IO interoperability), client (e.g. error concealment, behavior with packet loss and jitter), as well as RTP information (e.g. packet loss, jitter) and reference speech based information (e.g. location of packet loss, audio energy at that position based on Discontinuous Transmission, DTX, information). The ML features are then used to teach and optimize the ML algorithm towards the target value which is the MOS value (ITU-T P.863) measured immediately after the decoding, before the audio path. Thus, sQLEAR algorithm provides a network centric voice QoE predictor (MOS), free of the device's characteristics, such as automatic gain control (AGC), frequency shaping, voice enhancement devices (VED).

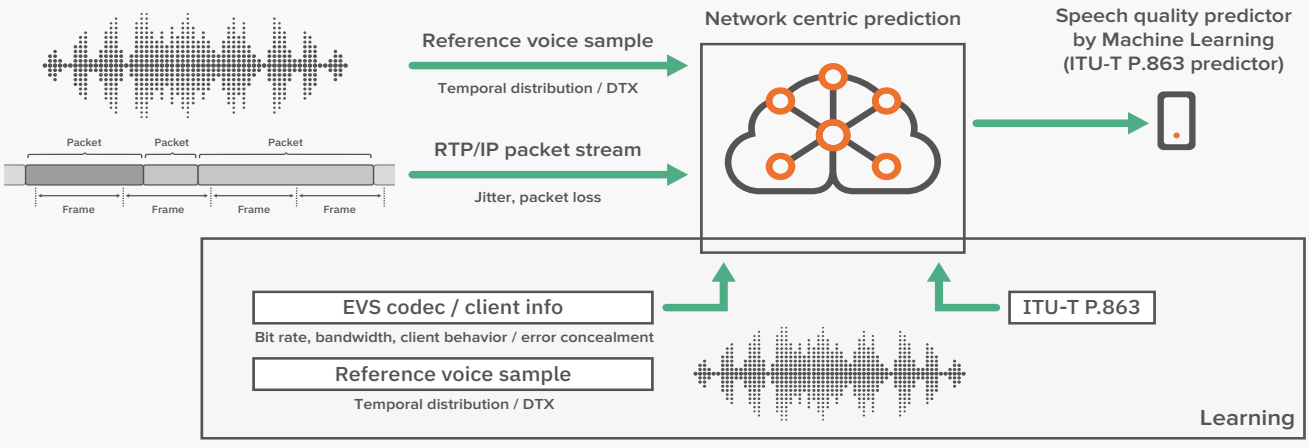


Figure 1. sQLEAR concept and operational mode.

sQLEAR is using a combination of two ML regression algorithms, Random Forest (RF) regressor and Support Vector Regressor (SVR) in order to ensure not only a high accuracy, but also a controllable and consistent accuracy. The two regressors (Figure 2 a,b) use different techniques to predict sQLEAR score. Random Forest algorithm (Figure 2a) uses a multitude of decision trees to predict sQLEAR score as the mean prediction (regression) of the individual trees. The SVR (Figure 2b) predicts sQLEAR as a function f(x) that deviates from the target values y (a.k.a ITU-T P.863 score) by a value not greater than ε for each test sample (a.k.a degraded speech file). For the test samples for which the constrain cannot be met, SVR uses two slack variables ξn and ξ*n * for each test sample, which allow regression errors to exist up to the value of ξn and/or ξ*n *, and still satisfy the required conditions. By keeping the predicted sQLEAR scores within the boundaries of the two slack values, allows SVR algorithm to act as a predictor corrector, as follows. It is expected that RF and SV regressors predict closely the same sQLEAR scores. If the difference between RF and SVR sQLEAR predictors is large, then it means that the difference goes outside the boundaries defined by the two SVR slack values. A decision function based on the FER value is used to decide which of the RF and SV regressors give the best result. Consequently, sQLEAR accuracy is controlled and kept consistent during operation/running time. More than 200000 samples have been used for learning and validation which resulted in performance predictions of 96%-98% correlation to ITU-T P.863. More about sQLEAR performance is discussed later on in the paper.

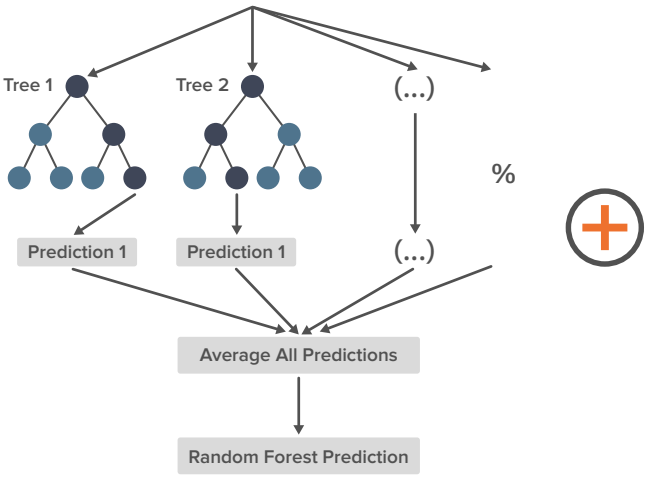


Figure 2a. Random Forest ML.

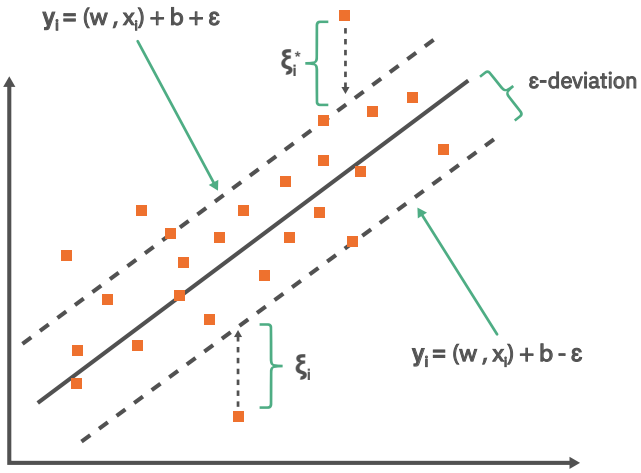


Figure 2b. Support Vector Regressor ML.

Figure 2. sQLEAR ML algorithms.

Operational mode

During the operational phase, sQLEAR measurement procedure is the same as for ITU-T P.863, in the sense that it sends a reference speech sample (same as used for learning/training and validation) to the system under test and predicts voice QoE from the combination of the output from the device and the sent reference sample (Figure 1). However, the output from the device differs between sQLEAR and ITU-T P.863. In the case of sQLEAR, the output from the device is the RTP stream (jitter, packet loss), while in case of ITU-T P.863, it is the recorded audio. If it is desired to perform additional analysis of quality concerning scenarios, the degraded speech file can still be recorded and stored for further analysis. The measurement set up scheme is presented in Figure 3.

The reference speech file is encoded by the device-based codec and sent over the transmission link on which it can be altered by various all-IP network error patterns (frame erasure FER, a.k.a packet loss, jitter). The encoding is performed for a specific bandwidth with different codec rates and modes (e.g. channel-aware mode, IO for EVS case). At the receiving end, the speech file is submitted to the EVS/AMR WB decoder and jitter buffer. The IP stream output is captured by a pre-processing module which decodes information regarding arrival time, sequence number and payload size, which is used to identify DTX presence, lengths and location, as well as to determine the codec bandwidth/ bit rate/mode used. The DTX information is used for synchronizing the reference speech file with the IP/RTP packet stream. The synchronization is achieved by correlating the DTX pattern and speech frames with the reference speech sample, which is stored at the receiving side. Through this process a binary file (.vqi) without FER (packet loss, jitter) during DTX periods (occurring during silence) is created to emulate the fact that errors present during silence do not impact the perceived speech quality. It should be noted that no recorded speech is needed to perform these tasks. The binary .vqi file along with the information regarding codec bandwidth/ bit rate/ mode is submitted to the module that calculates the ML features which are used as input to the pre-trained ML model. The output of the ML model is sQLEAR predicted voice QoE (a.k.a MOS).

The measurement set-up is mainly the same for all three use cases: VoLTE EVS, VoLTE AMR-WB and OTT/WhatsApp. The only difference comes with OTT voice. First, OTT voice use case does not use DTX, but Variable Bit Rate that has a similar behavior with DTX, resulting in much smaller packets during silence than during speech. Therefore, for OTT use case, these smaller packets play the exact role as the DTX packets for EVS/AMRWB use cases, as described above. Second, rather than using the OTT voice application- based client, a generic client is used. The main reason is the fact that the variety of OTT voice applications, with proprietary codecs and clients, and different levels of encryption become technically, practically and costly prohibitive to be tested in real live networks. Therefore, in order to establish a good benchmark, the generic client uses open source codec and client which are common among OTT voice services (e.g. OPUS codec, PJISP client). Open source codec/client offer many configurable settings which can be used to emulate very popular OTT voice applications (e.g. WhatsApp). More about this in new white paper to come on generic OTT testing.

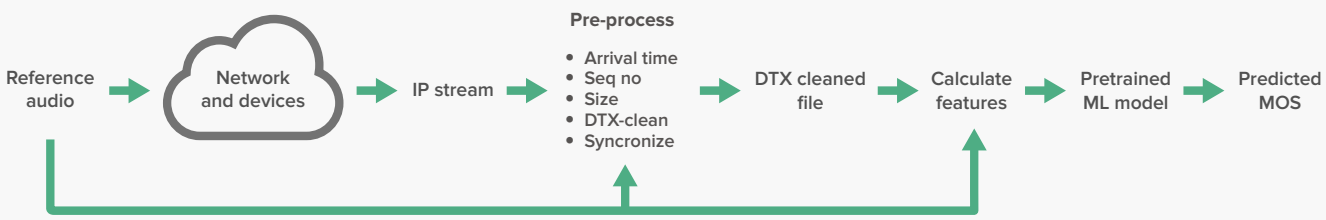


Figure 3. sQLEAR measurement set-up.

Which are the benefits of a network centric approach

Cost efficiency is today crucial for MNOs to manage and operate their 4G/LTE networks while deploying 5G. sQLEAR offers an optimal solution to this challenge by providing a network-centric voice QoE predictor, as described above. Thus, sQLEAR is free of device's impact (e.g. frequency shaping, AGC, VED), which comes with three important benefits. First, by using sQLEAR, a cost -efficient troubleshooting and/or optimization towards network issues rather than device issues can be performed. Second, independency of the device's audio path and any speech frequency shaping and/or processing within the device, ensures consistent and comparable sQLEAR scoring between different devices models. In addition, device transparency eliminates the need of tuning/calibration of the voice QoE predictor for each device model, and consequently enabling a prediction free of audio path artifacts. Third, the independency of the speech frequencies and content, and the network centeredness make languages less relevant for sQLEAR usability and operation.

In conclusion, network centric and device transparent approach makes sQLEAR a powerful voice QoE prediction solution for drive test scenarios designed for network troubleshooting and optimization, monitoring and regression testing, and significantly important for benchmarking campaigns when it is desired to remove the device's impact.

What should you know about machine learning based voice quality predictors

The unique machine learning techniques offer three clear benefits. First, machine learning algorithms are best suited to describe the continuously increasing complexity of the inter-dependencies between all network/codec/client parameters, as well as their impact and contribution to speech quality. Second, machine learning techniques are flexible and easy to tune to any changes that emerge from the introduction of new codecs/clients which a QoE predictor needs to account for; thus, enabling operational efficiency for QoE solutions using ML algorithms. Third, there is no need for additional calibration to the MOS scale using first or third order polynomials, because the machine learning based algorithm "learns" the precise MOS scale that it needs to predict. Thus, ML techniques are free of any accuracy artifacts that calibration procedures may involve.

However, these benefits are very sensitive to the possible misusage of the ML techniques; respectively, overfitting and underfitting of the ML algorithm towards the learning/training and validation data bases. Overfitting describes the bias of an algorithm towards the learning/training data sets, and underfitting defines a poor accuracy algorithm on leaning/training and validation data sets.

In this section the sQLEAR ML performance is analyzed for VoLTE EVS and AMRWB use case.

The process of ML learning/training from a data set involves the optimization of the ML algorithm's parameters, called "hyper parameters". The optimization is performed towards the best model, respectively highest accuracy (or minimal underfitting) on learning and validation data sets, and minimal, or preferably no bias, on the learning data set (least overfitting). The evaluation of these effects is performed using the Learning Curves technique [Jan N van Rijn et al. "Fast algorithm selection using learning curves"; Springer. 2015, pp. 298–309], which analyzes the ML algorithm's mean square error (MSE) for the learning/training

and validation data sets. An ML algorithm is not underfitted if it shows low learning/training MSE. An ML algorithm is not overfitted if the validation MSE continuously decreases with the increase of the size of the learning/training data set and finally reaches convergence to a minimum difference against the learning/training MSE. The Learning Curves technique is used regardless of the ML algorithm category; however, with slightly different nuances if supervised or unsupervised learning is used.

sQLEAR uses supervised ML from the regression algorithms' category, respectively the combination of RF regressor and SVR, as described above, and consequently benefiting sQLEAR with controlled and consistent accuracy. The results of the sQLEAR ML algorithms' performance is presented in Figure 4a (VoLTE EVS) and Figure 4b (VoLTE AMRWB). sQLEAR used a data set of about 130000 samples for VoLTE EVS and of about 30000 samples for VoLTE AMRWB; with a 50-50 split between learning/training and validation/testing set. Figure 4a,b shows that the ML algorithm used by sQLEAR has a very low MSE value on the learning/training data set, which proves that ML algorithm in sQLEAR is not underfitted or in other words is highly accurate, in both use cases. Figure 4a,b shows that the validation MSE converges fast to the learning/training MSE, for learning set size larger than about 70000 samples in EVS case (Figure 4a) and larger than about 15000 samples in AMRWB case (Figure 4b); respectively about the size used for learning data sets in each use case; sQLEAR has been trained on about 65000 samples for EVS use case and about 17000 samples for AMRWB sQLEAR. In addition, the trend of the validation error, in both use cases, does not deviate a lot and has steady decrease, which proves that having larger learning/training data sets cannot decrease the error significantly. Therefore, it can be concluded that ML algorithm used in sQLEAR is not overfitted.

In conclusion, ML based solutions demand using Learning Curves techniques to evaluate the ML algorithm's performance for overfitting and underfitting. The results presented and discussed in this section (Figure 4a,b), show that the ML algorithm used in sQLEAR is free of overfitting (bias towards a data set) and underfitting (poor accuracy), for both VoLTE EVS and AMRWB use cases. More on sQLEAR overall performance and accuracy is discussed in the next section.

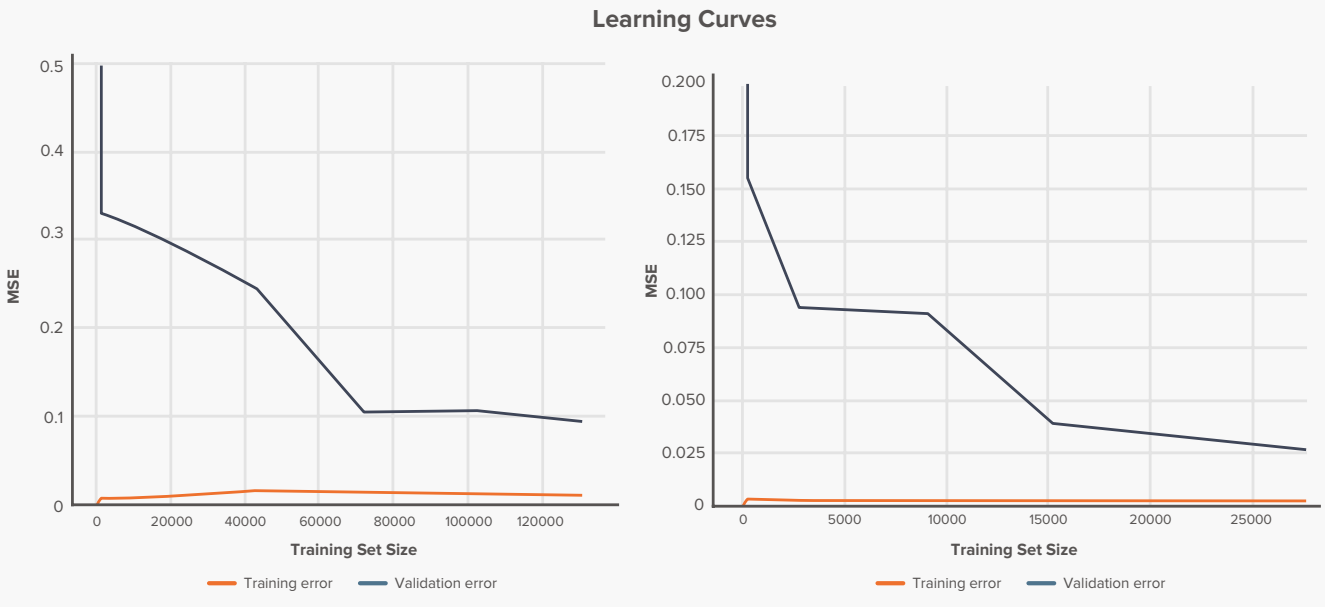


Figure 4a. VoLTE EVS.

Figure 4b. VoLTE AMRWB.

Figure 4. sQLEAR - ML overfitting / underfitting results

Which are sQLEAR lab and field trials validation results

The thorough and complete sQLEAR validation involved three level of testing.

Tests performed based on simulated device using an audio path without degradations. This is in alignment with ITU-T P.565 and sQLEAR performance must meet minimum requirements as defined in the framework. Results of these tests are discussed above.

Tests with real devices showing good audio paths. In this case sQLEAR and ITU-T P.863 are running simultaneously using same type of device under controlled conditions, which are required to cover the entire quality scale, from poor to very good radio coverage, in order to ensure meaningful analysis (e.g. correlation coefficient calculation). Results are discussed and presented above.

Tests performed by customers as independent parties. These are drive test scenarios comparing sQLEAR and ITU-T P.863 on a variety of devices and running in various radio environments. Results and conclusions on these tests are presented above.

Minimum performance requirements as defined by ITU-T P.565

sQLEAR is developed based on ITU-T P.565 framework, and consequently its performance has been evaluated against the framework's minimum performance requirements on the ITU-T provided test vectors, learning and validation data sets, and reference speech sample. As required by ITU-T, the validation run in two phases: first with "known" simulated validation data and second with "unknown" live validation data. By "known" is meant that the data has been used during the ML learning-validation phases; by "unknown" is meant that the voice QoE predictor has never been exposed to this data. Both these validation data sets are per P.565/Annex F.

VoLTE EVS use cases

The sQLEAR results on the unknown live validation data set for VoLTE -EVS use cases are presented in Figure 5. Figure 5a shows for sQLEAR a correlation coefficient of 0.98, an RMSE (root mean square error) of 0.27MOS and a mean absolute error MAE (mean absolute error) of 0.17MOS, when compared with ITU-T P.863. All these performance metrics meet the ITU-T P.565 minimum requirements which are: correlation coefficient less than 0.96, RMSE/MAE <0.3MOS (P.565/Annex D). Figure 5b shows the distribution of the prediction error (a.k.a absolute error, AE) as the absolute difference between sQLEAR score and ITU-T P.863 score per individual speech sample. In this case as well, sQLEAR meets the minimum performance requirements (see Table 1, EVS).

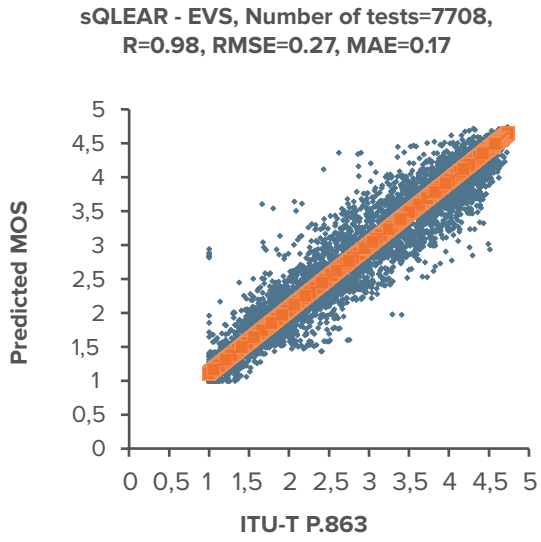


Figure 5a. Correlation performance metric.

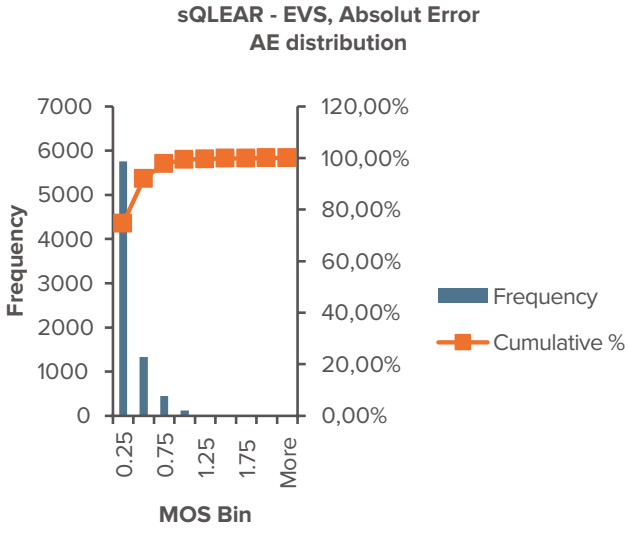


Figure 5b. Absolute error performance metric.

AE		MOS < 0.25	MOS < 0.5	MOS < 1
ITU-T P.565 Requirements		70%	90%	99%
sQLEAR	VoLTE EVS	74.72%	91.98%	99.36 %
	VoLTE AMRWB	81.59%	95.17%	99.8%

Table 1. sQLEAR prediction error distribution for VoLTE EVS and AMRWB unknown live validation data set (ITU-T P.565).

Intrusive parametric voice QoE predictors such as sQLEAR are mainly designed, and consequently most suitable, for drive testing scenarios. Therefore, from a drive testing perspective, it is crucial that solutions such as sQLEAR accurately predict poor quality as well as correctly differentiate quality levels and identify quality trends and variations. While accurate prediction of poor quality is important for network troubleshooting and optimization, the correct estimation of the high quality is equally important because it ensures that optimization effort is not spent when/where there is no need.

In addition, it is required a high accuracy in order to be able to differentiate the quality differences on the upper end of the scale. This is needed since even small differences can create a big impact if there are a lot of values at the upper end of the quality scale. Figure 6a,b shows sQLEAR performance in a broad range of FER conditions, for various encoding rates, both EVS as well as EVS AMR-IO mode. sQLEAR follows closely ITU-T P.863 behavior and trend on the whole range of conditions from very good (low FER values) to very poor quality (high FER values), with almost no differences in very poor conditions.

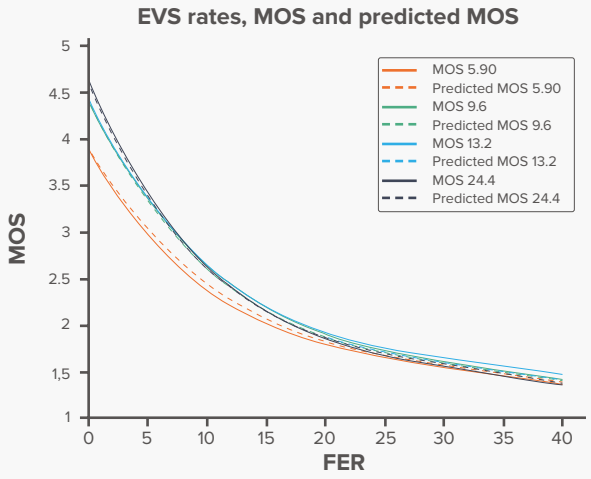


Figure 6a. EVS bit rates.

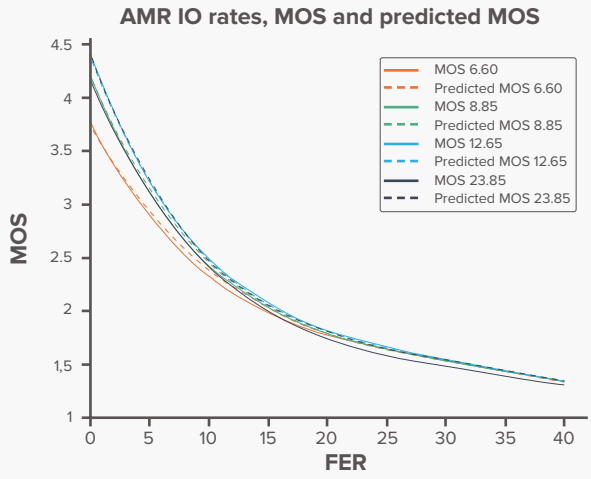


Figure 6b. EVS AMR-IO mode.

Figure 6. sQLEAR and ITU-T P.863 vs FER at different encoding bit rates, EVS use case.

Table 2 (EVS/AMR-IO) presents sQLEAR performance when predicting the highest quality for each EVS/EVS-AMR-IO rates. sQLEAR either shows very small differences, and only at the second decimal, or no differences for some of the bit rates, when compared to ITU-T P.863. This is in agreement with ITU-T P.565 – Appendix IV.

Codec	Bit rates	Max MOS (ITU-T P.863)	Max Predicted MOS (sQLEAR)	Difference
EVS, AMR-IO	5.90	4.06	4.08	-0.02
	6.60	3.93	3.91	0.02
	7.20	4.16	4.18	-0.02
	8.00	4.27	4.22	0.05
	8.85	4.36	4.33	0.02
	9.60	4.55	4.54	0.01
	12.65	4.55	4.53	0.02
	13.20	4.57	4.57	0
	16.60	4.71	4.71	0
	23.85	4.58	4.57	0.01
	24.40	4.74	4.74	0
AMRWB	6.60	3.63	3.60	0.03
	8.85	4.04	4.02	0.02
	12.65	4.26	4.25	0.01
	23.85	4.41	4.38	0.03

Table 2. ITU-T P.863 and sQLEAR for EVS and AMRWB rates.

VoLTE AMR-WB use case

sQLEAR results on the unknown live validation data set for VoLTE -AMR-WB show similar results as for EVS use case. Figure 7a shows a correlation coefficient of 0.98, an RMSE of 0.22MOS and a mean absolute error MAE of 0.14MOS, when compared with ITU-T P.863. As in EVS case, all three-performance metrics meet the ITU-T P.565 minimum requirements. Figure 7b shows the distribution of the prediction error (AE) which also meets the minimum performance requirements (see Table 1, AMRWB).

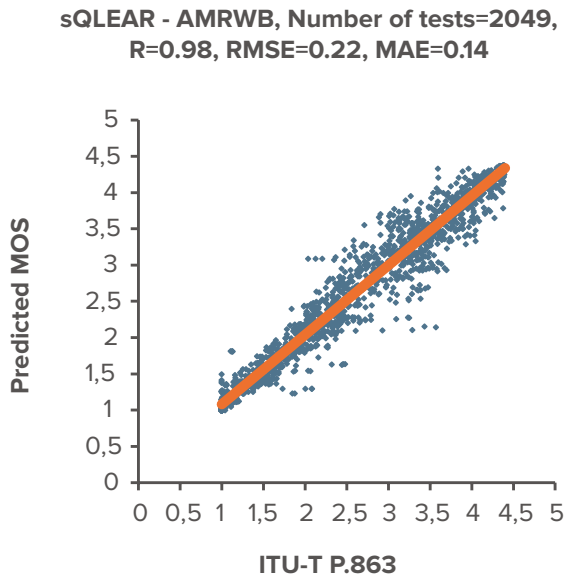


Figure 7a. Correlation performance metric.

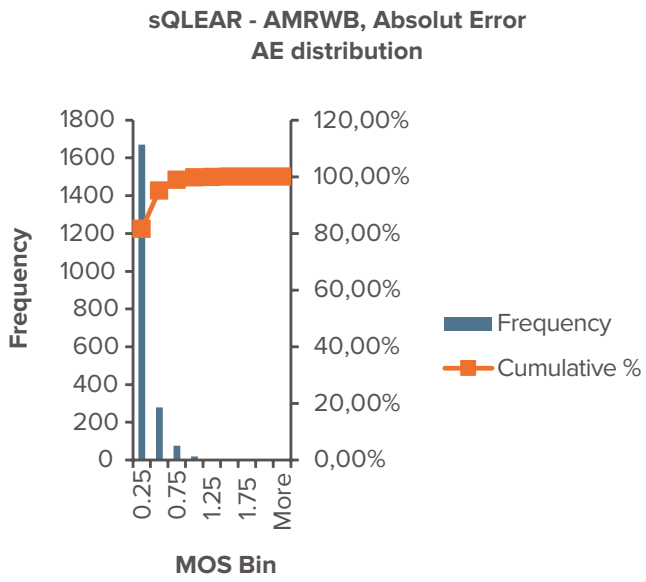


Figure 7b. Absolute error performance metric.

Like for EVS use case, AMRWB case (Figure 8) shows that sQLEAR follows closely ITU-T P.863 behavior and trend on the whole range of network conditions from very good (low FER values) to very poor quality (high FER values), with almost no differences in very poor conditions.

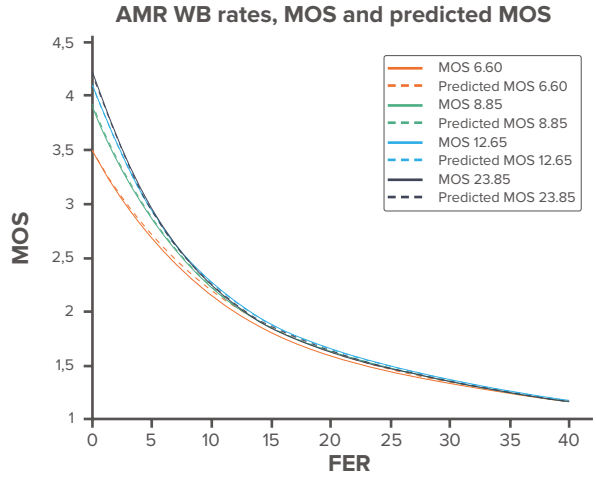


Figure 8. sQLEAR and ITU-T P.863 vs FER at different encoding bit rates, AMRWB use case.

Like in EVS use case, sQLEAR predictions of the highest quality for AMRWB rates show either very small differences, and only at the second decimal, or no differences for some of the bit rates, when compared with ITU-T P.863 (Table 2, AMRWB); thus, in agreement with ITU-T P.565 – Appendix IV.

In conclusion, sQLEAR meets ITU-T P.565 minimum performance requirements, for both VoLTE EVS and AMRWB use cases.

Validation in lab and field trials

In order to validate sQLEAR robustness and performance consistency, additional validation tests in lab and field trials have been performed with several completely new ("unknown") data sets.

Overall performance

These tests used real devices with different audio paths, but most devices using digital audio and extensively tuned to ensure that ITU-T P.863 scores are not significantly affected by the audio path. Taking care of these aspects is crucial to ensure a meaningful comparison between sQLEAR and ITU-T P.863.

The test conditions covered the whole quality range to ensure meaningful correlation calculations and used various devices, both EVS and AMRWB based. For VoLTE EVS the following devices have been used: Samsung 960F (9.9kb/s, 24.2kb/s), Sony XZ2 (9.6kb/s), F8141 (13.2kb/s). For VoLTE AMRWB the following devices have been used: Samsung 960U (12.65kb/s, 23.85kb/s), and 977N (23.85kb/s), Sony XZ2 (23.85kb/s).

During the trials, the devices have been locked on each codec at a time, and on various bit rates, in order to run validation analysis per individual use case. In addition, for each use case, sQLEAR and ITU-T P.863 run simultaneously, using same device type/model.

Correlation, RMSE, MAE and AE distribution has been calculated for EVS (about 1800 samples) and AMRWB (about 1100 samples) data sets. The validation results are presented in Figure 9, Figure 10 and Table 3 and they show for both use cases that although these data sets are completely new to sQLEAR and contain various live network conditions, sQLEAR exhibits high performance on all metrics, correlation coefficients of about 98%, RMSE and MAE lees than 0.3MOS.

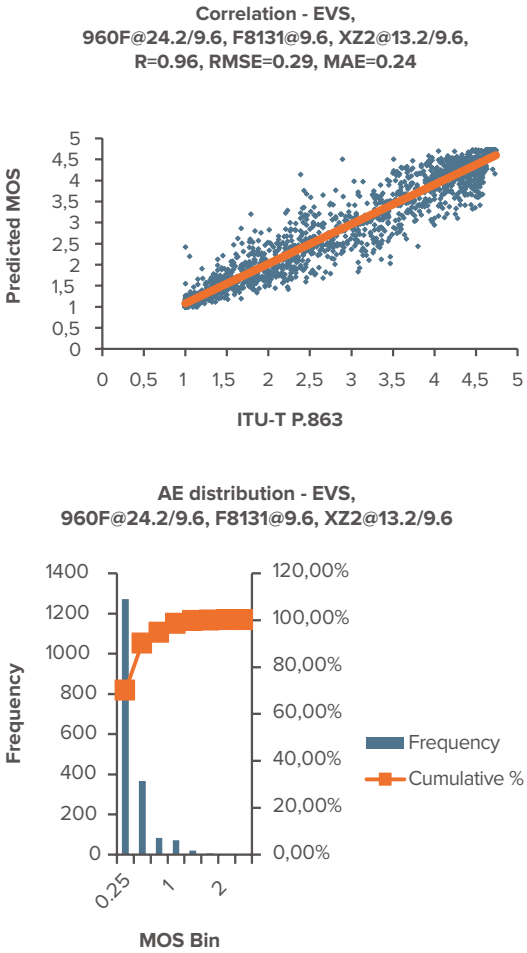


Figure 9. sQLEAR performance results for VoLTE EVS use case, lab and field data

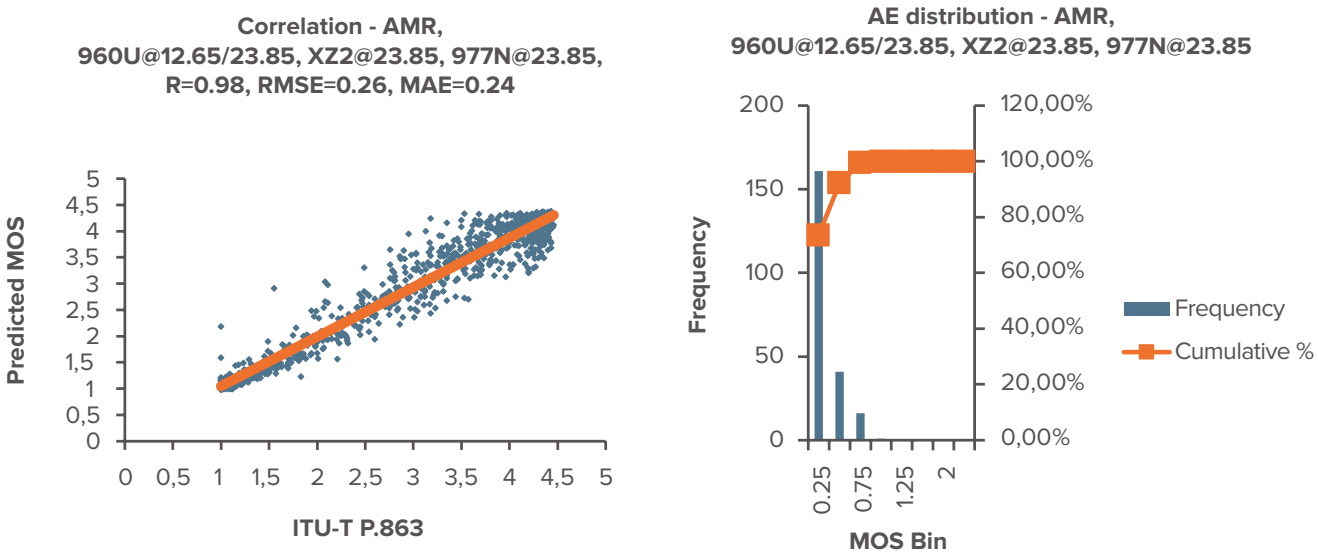


Figure 10. sQLEAR performance results for VoLTE AMRWB use case, lab and field data.

AE		MOS < 0.25	MOS < 0.5	MOS < 1
sQLEAR	VoLTE EVS	70.01%	90.1%	99%
	VoLTE AMRWB	73.52%	92.24%	100%

Table 3. sQLEAR prediction error distribution for VoLTE EVS and AMRWB lab and field trials data.

In addition, sQLEAR scores distribution versus ITU-T P.863 scores have been analyzed and validated. The results (Figure 11) show that the two distributions, sQLEAR and ITU-T P.863 scores, are statistically significant close to each other. More importantly, results show that sQLEAR distinguishes in agreement with ITU-T P.863 both peak poor quality as well as peak good quality.

In conclusion, sQLEAR can accurately detect quality trends, which makes sQLEAR a reliable voice QoE predictor for monitoring, troubleshooting and optimization towards network issues, as well as benchmarking.

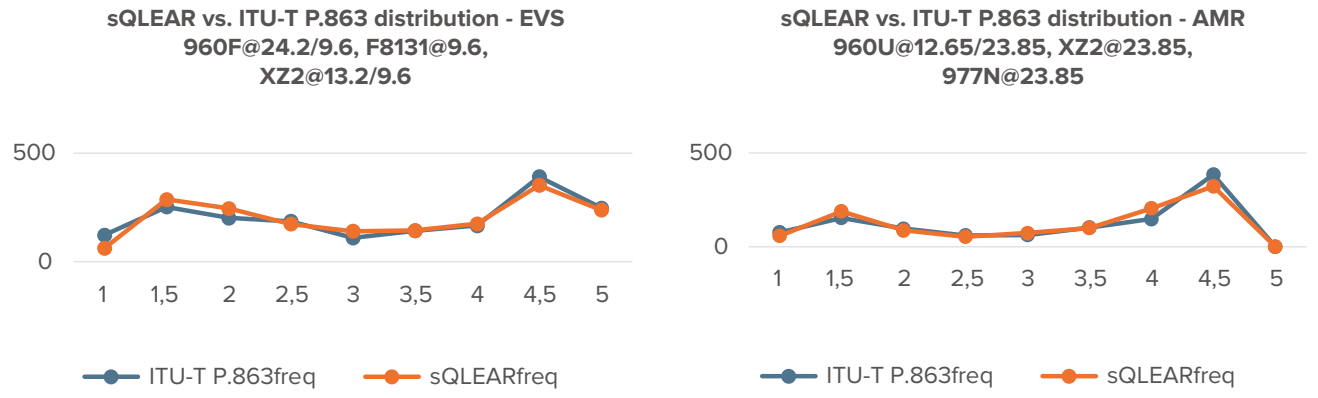


Figure 11a. VoLTE EVS.

Figure 11b. VoLTE AMRWB.

Figure 11. sQLEAR vs. ITU-T P.863 scores distributions.

Consistency

As a network centric QoE predictor, sQLEAR is expected to show consistent results across devices. Different devices, using the same codec rate and running in the same network conditions should show the same sQLEAR score. Since it is difficult to define the exact same network conditions, the maximum achieved sQLEAR score across all the devices has been analyzed. The results (Table 4) show that for all the devices, per each codec rate, the maximum sQLEAR score achieved is the same to the first decimal. Thus, sQLEAR shows consistency across devices.

Device@rate	VoLTE AMR			VoLTE EVS	
sQLEARmax	960U@12.65	960U@23.85	XZ2@23.85	960F@9.6	F8131@9.6
	4.33	4.35	4.35	4.53	4.53

Table 4. sQLEAR consistency across devices.

Exploiting the benefits of device independency

During the validation process, detailed analysis on sQLEAR and ITU-T P.863 scores has been performed for each device-codec rate combination. In the case of Samsung 960F / VoLTE AMRWB, 12.65kB/s and 23.85kb/s, the sQLEAR scores distribution showed an unexpected shift when compared with ITU-T P.863 scores (Figure 12); with sQLEAR unveiling more optimistic voice quality. The phenomenon has been detected only on one device, AMRWB codec, and further investigations showed that the processing on the audio path of the device was the reason for artificially degrading ITU-T P.863 scores. Consequently, using such a device presents the risk of covering real problems in the network, and therefore, sQLEAR is preferred over ITU-T P.863 in these cases. Free of device's impact, sQLEAR proves to be able to certify devices which can be safely used for perceptual voice quality prediction.

However, it should be noted that if analog audio is used for the tests, then always a certain level of degradation of the audio path is expected and consequently artificially impacting ITU-T P.863 scores, but not sQLEAR scores which predict the network centric quality, free of the audio path impact.

In conclusion, extensive validation tests with various lab and field conditions, from poor to high quality, and different devices and codec rates combinations, show that sQLEAR exhibits high performance. The validation results also prove that sQLEAR can be safely used for monitoring, troubleshooting, optimization and benchmarking. In addition, sQLEAR consistency across devices has been validated, which turns out to be a safe method to certify the devices which are appropriate to be used with ITU-T P.863 without the danger of introducing the artifact of a poor audio path.

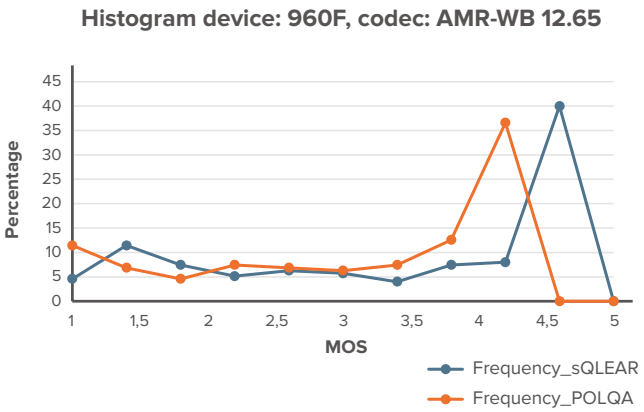


Figure 12. Detecting poor device audio path using sQLEAR.

What does an independent sQLEAR validation unveil

Validation results presented above proved that sQLEAR predicts accurately voice QoE both on ITU-T P.565 test samples as well as on a large set of unknown lab and field collected data.

In addition, for a thorough and complete validation, sQLEAR has been also submitted to an independent validation performed by an Infovista customer. The independent validation run sQLEAR and ITU-T P.863 simultaneously, using various devices in a variety of stationery and drive test scenarios and findings showed nothing less than what Infovista expected. sQLEAR "shows small difference between quality scores for various types of devices and the impact of the device type itself on the scores is significantly smaller than using ITU-T P.863", a desired effect when optimizing and troubleshooting voice services' quality. In addition, "like ITU-T P.863, sQLEAR correctly changes to low values in areas of poor RF quality", which proves sQLEAR accurate sensitivity to quality degradations, also a crucial feature of a voice QoE predictor like sQLEAR for optimization and troubleshooting. Last, but not least, the independent party field trials showed the same distribution of sQLEAR and ITU-T P.863 scores, with sQLEAR "results close to ITU-T P.863 vs.3 suited for EVS FB", which proves sQLEAR suitability for high definition voice quality prediction as well as its readiness for 5G voice services' quality evaluation. These observations are reflected in the performance results presented in Figure 13a (EVS case) and Figure 13b (AMRWB case). The results are based on validation tests using 2 pairs of Samsung S10 devices making mobile to mobile calls; one pair for sQLEAR and the other one for ITU-T P.863. The charts show the similarity of the distributions, the same best quality bin as well as the same sensitivity to poor quality degradations. sQLEAR network centric characteristic, and its capability of being device independent, is reflected in the small differences shown in the distributions.

In conclusion, based on all validation tests, the independent party stated that "test results give the opportunity to use sQLEAR for accurate quality monitoring measurements and benchmarking campaigns".

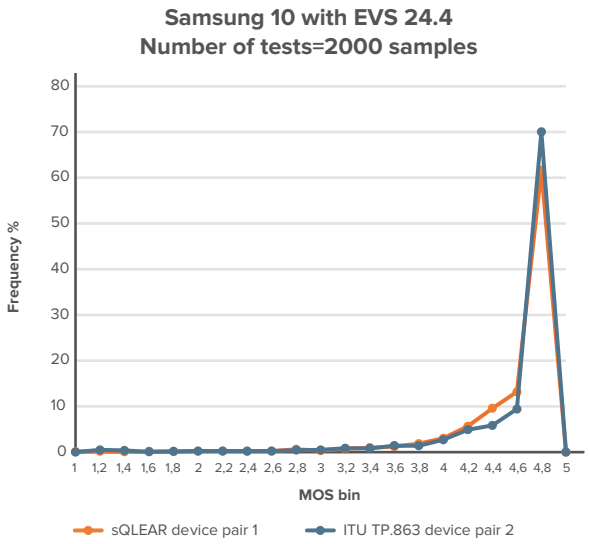


Figure 13a. sQLEAR and ITU-T P.863 with live data (Samsung 10, EVS 24.4kb/s).

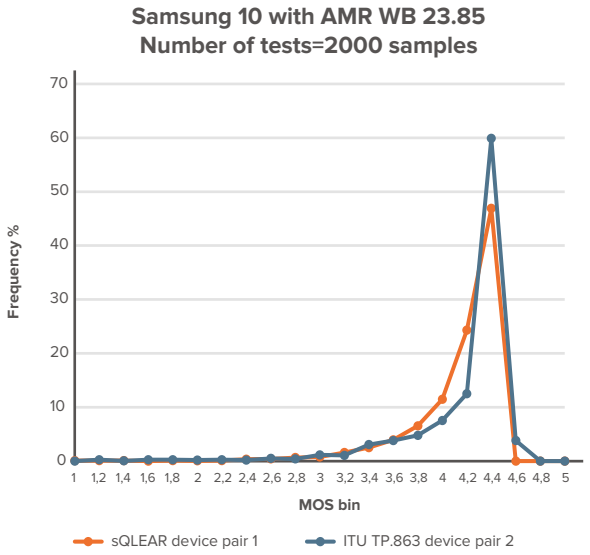


Figure 13b. sQLEAR and ITU-T P.863 with live data (Samsung 10, AMRWB 23.85kb/s).

Which are sQLEAR supported test senarios in Infovista portfolio

Designed and developed for voice services over all-IP mobile networks, 4G/LTE and 5G, sQLEAR aims to support both carrier (e.g. VoLTE, VoNR) and OTT (e.g. WhatsApp) voice solutions. sQLEAR can predict HD VoLTE EVS (with IO and CA modes) and AMR-WB voice, and it is ready be used for HD VoNR EVS, once the 5G voice service is available in the network. In addition, sQLEAR aims to predict voice QoE of OTT applications using generic OTT client emulating WhatsApp service behavior. Table 5 shows sQLEAR released test scenarios in Infovista portfolio.

	Infovista Products	
	TEMS™ Investigation 22.0	TEMS™ Paragon 4.0
Test Scenarios	VoLTE EVS , EVS AMR-IO	VoLTE EVS, EVS AMR-IO
	VoLTE AMR WB	VoLTE AMR WB

Table 5. sQLEAR consistency across devices.

As a network centric metric, free of device's impact on the voice quality, sQLEAR is best suitable for drive testing scenarios aimed for troubleshooting and optimization, monitoring and regression testing with TEMS™ Investigation and benchmarking with TEMS™ Paragon.

How should one set up an independent sQLEAR validation with TEMS™

As any machine learning based solution, sQLEAR can benefit of extensive independent validations, especially in live networks scenarios which can in the long run enrich sQLEAR learning of behavior of a large variety of possible network equipment configurations. Therefore, besides the fact that they prove sQLEAR performance, independent validations can eventually enable further enhancements for next sQLEAR releases. The latter becomes crucial for network testing tools in today's dynamically evolving networks; and sQLEAR machine learning technique is best suited to support this.

However, unless carefully and correctly set up, an independent validation can provide inaccurate and misleading results. Therefore, the following best practices should be followed if an independent validation of sQLEAR versus ITU-T P.863 is set up with TEMS™ Investigation.

- **Drive test route:**
  Select a route which covers from poor to very good network conditions in order to make sure that the MOS voice quality scale is evenly covered, and consequently it is ensured that accurate and unbiased correlations can be performed on the collected data (sQLEAR and ITU-T P.863 scores). Drive the route several times and for each time switch which devices are running sQLEAR/ITU-T P.863 to ensure that there is no offset from the position of the devices in the car.
- **Equipment set up:** Set up two laptops (one for sQLEAR, one for ITU-T P.863) to run mobile to mobile (M2M) calls, Device Digital Audio, ITU-T P.863 SWB/sQLEAR using the following test set up configuration:
  - Language: English 3 (US) (per ITU-T P.565/ AnnexF)
  - Set AudioOption=11 in QualityMeasurementConfig to store all trace files from sQLEAR for the analysis
  - In Voice Quality activity set "Store AQM Files: Yes" and "MOS limit: 5" in order to store both the audio files recorded during sQLEAR (not used by sQLEAR) and the .vqi files which contain the RTP stream used by sQLEAR. In this way it is ensured that in the eventuality of any needed verification and/or double checking, then ITU-T P.863 can be run offline on the recordings corresponding to the .vqi files.
  - Lock on 4G to make sure that only VoLTE to VoLTE calls are set-up
  - Lock to the codec and the codec rate in order to enable a controlled validation of the use cases. (e.g. suggestion: EVS 9.6 and AMR WB 12.65). It should be noted that same codec and codec rate must be set up on both sQLEAR and ITU-T P.863.

- **Devices**
  - Use four identical devices from the following certified list; by certified it is meant that all show a good audio path so that sQLEAR to ITU-T P.863 comparison is free of test equipment artifacts: 973F, 977F/N, 960U, Sony XZ2.
  - It should be noted that this list refers only if a validation against ITU-T P.863 is desired; otherwise, devices with poor audio path can be used with sQLEAR since the algorithm is not using the audio path, as already mentioned.
  - Use the same SIM and network operator for all devices in order to avoid possible transcoding that degrade ITU-T P.863 score but not the sQLEAR score.
- **Validation analysis**
  - Export the sQLEAR and ITU-T P.863 scores to Excel
  - Evaluate outliers (e.g. random and sparsely distributed samples showing more than 0.75MOS prediction errors) and use the samples for more detailed further analysis (e.g. detection of devices with faulty audio path).
  - Calculate regressions, histograms (0.25MOS bins), average and standard deviations.
    It should be noted that if the test samples do not cover evenly the entire MOS scale, then correlation coefficient is invalid.

What should one learn about sQLEAR

Infovista machine learning based sQLEAR algorithm predicts voice QoE using IP transport, codec and jitter buffer in the end-user voice client information, and the temporal speech distribution within the reference sample, without the need of recording the resulting degraded speech sample. Consequently, sQLEAR predicts the network centric view of the voice QoE independently of any speech frequency shaping and/or speech processing within the device. Therefore, free of device's audio path impact, sQLEAR enables cost effective, network centric monitoring, optimization and troubleshooting and benchmarking of the EVS (CA, AMR-IO) and AMR WB based VoLTE QoE. As a network centric and free of device's impact solution, sQLEAR is best suitable for Infovista drive testing portfolio, such as TEMS™ Investigation and TEMS™ Paragon.

sQLEAR has been developed based on ITU-T P.565 framework and it proves to meet the ITU-T framework's minimum performance requirements, on all the data bases provided by ITU-T. In addition, sQLEAR validation results in live network conditions show that it accurately predicts voice QoE by closely following ITU-T P.863 scores' distribution and trends. Infovista customers performed independent validations and based on the results they concluded that "test results give the opportunity to use sQLEAR for accurate quality monitoring measurements and benchmarking campaigns."

KNOW YOUR NETWORK™

About Infovista

Infovista, the leader in modern network performance, provides complete visibility and unprecedented control to deliver brilliant experiences and maximum value with your network and applications. At the core of our approach are data and analytics, to give you real-time insights and make critical business decisions. Infovista offers a comprehensive line of solutions from radio network to enterprise to device throughout the lifecycle of your network. No other provider has this completeness of vision. Network operators worldwide depend on Infovista to deliver on the potential of their networks and applications to exceed user expectations every day. Know your network with Infovista.