NIST Technical Note 2171

Optimal Transmit Volume Conditions for Mission Critical Voice Quality of Experience Measurement Systems

Chelsea Greene
Jesse Frey
William Magrogan
Cara O'Malley
Jaden Pieper

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2171

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Technical Note 2171

Optimal Transmit Volume Conditions for Mission Critical Voice Quality of Experience Measurement Systems

Chelsea Greene
Jesse Frey
William Magrogan
Cara O'Malley
Jaden Pieper
Public Safety Communications Research Division
Communications Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2171

September 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology Technical Note 2171
Natl. Inst. Stand. Technol. Tech. Note 2171, 34 pages (September 2021)
CODEN: NTNOEF

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2171

Acknowledgments

The authors would like to acknowledge Hossein Zarrini for his contributions to preliminary data collection and distortion measurement design. The authors would like to acknowledge Steve Voran for his wealth of knowledge on audio quality and willingness to discuss new ideas and pitfalls. The team would also like to thank Don Bradshaw for his vocoder and audio signal insight.

Abstract

Selecting transmit volume levels using a transmit volume optimization (TVO) process is the first calibration step in performing quality of experience (QoE) measurements developed by the Mission Critical Voice portfolio of the Public Safety Communications Research (PSCR) Division. As noted in prior publications, audio volume levels have an impact on output consistency while performing mission critical voice (MCV) QoE measurements with push-to-talk (PTT) communications devices. These measurements must be consistent, repeatable, and comparable. The goal of this project is to ensure that MCV measurements are consistent and repeatable; the project focuses on transmit volume levels and their impact on the system under test (SUT). A measurement that characterizes audio distortion levels, specifically caused by overdriven speech into a transmit device, was developed. This measurement is performed across a series of transmit volume levels. The results are used to present the system user with optimal transmit volume levels to ensure MCV measurements contain a minimal amount of uncertainty caused by distortion within a range of stable volume levels. This paper will discuss the development of this specific distortion measurement and the methods designed to find optimal transmit volume levels.

Key words

Analog; Audio; A-weight; Communications; Direct mode; Distortion; Frequency; Frequency slope fit (FSF); Key performance indicator (KPI); Land mobile radio (LMR); Mission critical push-to-talk (MCPTT); Mission critical voice (MCV); Mouth-to-ear (M2E); Optimal volume plateau identification algorithm (OVPIA); Project 25 (P25); Plateau; Public safety; Push-to-talk (PTT); Quality of experience (QoE); System under test (SUT); Transmit volume optimization (TVO); Volume.

Table of Contents

List of Tables

List of Figures

Acronyms

FSF frequency slope fit. i, 2–14, 18, 20, 26

KPI key performance indicator. i, 1

LMR land mobile radio. i

M2E mouth-to-ear. i, 1, 2, 8, 22

MCPTT mission critical push-to-talk. i, 1, 6

MCV mission critical voice. i, 1–3, 8, 15, 16, 20, 22

MRT Modified Rhyme Test. 8

OVPIA optimal volume plateau identification algorithm. i, 2, 3, 6–8, 12–17, 20

P25 Project 25. i, 8, 9, 11, 13, 15–17, 25, 26

PESQ perceptual evaluation of speech quality. 9, 12, 14–20

PSCR Public Safety Communications Research. i, 1, 16, 20

PTT push-to-talk. i, 1–3, 8, 10, 13, 20, 22

QoE quality of experience. i, 1, 6, 8, 15, 16, 20, 22

SUT system under test. i, 1–4, 7, 8, 10–13, 15–18, 20

TVO transmit volume optimization. i, 1–3, 6, 8–10, 13, 15, 16, 18, 20, 24

Symbols

b Periodogram frequency bins. 4

D Difference in audio quality. 14, 15

L Lower limit of periodogram frequency bin. 4

m Slope from peak power bin through all higher frequency bins. 4, 5

N The number of trials for a technology. 14, 15

n Number of periodogram samples. 4

P PESQ score. 14

p Periodogram signal power. 4

U Upper limit of periodogram frequency bin. 4

V Volume level. 14, 15

$V_{\mathbf{Rx}}$ Receive volume level. 1, 2, 22

$V_{\mathbf{Tx}}$ Transmit volume level. 1–3, 6–10, 12–18, 20, 22

w Weighting constant. 15

1. Introduction

The National Institute of Standards and Technology's (NIST) Public Safety Communications Research (PSCR) Division has developed multiple mission critical voice (MCV) quality of experience (QoE) measurements for mission critical push-to-talk (MCPTT) communications systems. It is essential to deliver high quality measurements for use by those vested in public safety communications. In this process, it is important to ensure that proper measurement controls are implemented for repeatable measurements. Volume level settings are an important test parameter for push-to-talk (PTT) devices within a system under test (SUT) and must remain consistent between measurements.

This project aims to develop a system that will classify audio distortion caused by overdriving transmitted speech and use that knowledge to present an optimal transmit volume, V_{Tx}, for users. The process of identifying this setting is referred to as transmit volume optimization (TVO). Establishing a method to identify proper volume settings is essential to high quality, repeatable, comparable, and interpretable measurement systems for MCV QoE key performance indicators (KPIs). This measurement provides an essential calibration step to perform other MCV QoE measurements. This paper builds on the work of previous PSCR MCV measurement systems. Familiarity with the MCV measurement setup developed in prior publications, such as mouth-to-ear latency [1] and end-to-end access time [2], as well as the access time addendum [3], is highly recommended for full context of this body of work.

2. Background

Volume levels, both on transmitting and receiving MCPTT communications devices, can impact measurement quality. Here V_{Rx} is defined as the receiving user's device volume level. For a user, V_{Rx} impacts how well a user hears a message from a loudness perspective. V_{Rx} also impacts device performance, particularly if the audio chain of the MCPTT communications devices are overdriven. In the audio interface, there are a variety of device elements between the audio jack and the analog-to-digital converter, such as amplifiers. If the voltage into an analog-to-digital converter is greater than the full-scale voltage [4], distortion caused by over scaling is possible. Alternatively, an analog amplifier may be driven into its non-linear region. V_{Tx} is defined as the transmitting user's volume level. Transmitted audio that is loud and overdriven or too quiet to process can cause devices to compensate and process audio in ways not ideal to performing controlled QoE measurements. In practicality, one would not yell loudly or whisper into a PTT communications device; selecting a V_{Tx} within a range of safe values, away from edge cases, is ideal.

Ensuring consistent input into PTT communications devices for testing means devices behave the same way; when devices behave the same way under test, all technologies can be measured equally and comparably. During early mouth-to-ear (M2E) latency research [1], it was found that V_{Tx} affects results in ways that were statistically significant. Results were not repeatable or consistent, as the volume levels used were not maintained across measure-

ments or PTT devices. Thus, there needed to be a method to set the V_{Tx} in a standard way for any PTT communications device in a SUT. In a controlled testing environment, devices should maintain consistent behavior across multiple tests. By selecting volume levels that avoid clipping, PTT devices will function within their designed nominal ranges and avoid doing additional work to compensate.

The test setup used to perform MCV measurements contains a transmitter and a receiver PTT device. These devices function in the test setup with an audio interface as described in Sec. 4.3 of Ref. [2]. V_{Tx} is controlled via the audio interface settings used in the measurement system and is the focus of the work described in this paper. V_{Rx} is controlled on the communications device and the received audio gain knob on the audio interface; both of these receive volume levels remain the same throughout testing and are outside of the scope of this measurement. V_{Rx} is fairly straightforward to set due to the dynamic range in PTT communications devices that protects from reaching settings that cause overdriving.

A procedure to select an optimal V_{Tx} was developed and described in the M2E paper [1]. This procedure used an audio quality measurement and a peak finder to select an optimal V_{Tx}. The output of this test delivers the specific level to be used in MCV measurements performed using that particular PTT device and audio file. Throughout this paper, this procedure is referred to as the TVO. Updates to the TVO are the focus of this body of work. This internally developed measurement offers a calibration method that is inherently designed to work with the preexisting MCV measurement system to use before taking other measurements.

3. Measurement Design

The purpose of developing methods to ensure high quality MCV measurements is to minimize, within the SUT, factors that cause variation in results. Realistically, devices will not have one single optimal level where they operate in a standard mode. With this statement in mind, two assumptions were made. First, that an optimal range of volume levels exists and should be identified. This range can then be used as a basis for defining a single optimal V_{Tx}. Second, the concerns for the V_{Tx} mostly revolve around clipping-related distortion caused by overdriven audio. If the impact of this particular distortion could be measured, then this information could be used to establish the optimal V_{Tx} range.

Within the TVO, measurements are performed by varying V_{Tx}. A test is defined as running the TVO script one time to create one data set with an output consisting of an optimal V_{Tx}. Two elements are required for the TVO, using the two assumptions defined in the start of this section. The first measurement, frequency slope fit (FSF), was developed to measure distortion caused by a suboptimal V_{Tx}. A trial is defined as one iteration of the process of a PTT device transmitting audio through the measurement system and that audio being received by the receiver device. That trial produces one FSF score. The TVO script, by default, performs 40 trials at each evaluated V_{Tx}, 10 of each of the four talker's audio file described in Sec. 4.1. FSF is calculated for each trial, such that there is one score for every trial in a test. The second measurement, optimal volume plateau identification

algorithm (OVPIA), is used to find a range of optimal volume levels and select evaluation points. The detected range is used to select the optimal V_{Tx} to be used for all other MCV measurements for that particular SUT and audio files. The number of total trials per test will vary, as the number of V_{Tx} levels evaluated will depend on OVPIA. Typically, 16 to 20 levels are evaluated per run of the TVO.

3.1 Frequency Slope Fit

FSF is the first element of the TVO used to detect changes in behavior caused by an overdriven V_{Tx}. FSF characterizes the rate at which the high frequencies in the power spectrum fall off. A few steps are taken to lead up to this characterization. This measurement makes sense on a mathematical level; as a waveform is clipped, harmonics are produced, which results in more of the waveform power ending up at higher frequencies. Thus, as V_{Tx} approaches levels that induce clipping, the amount of power in the high-frequency bands increases, reducing the steepness of the slope. FSF is calculated as the change in slope of power across different frequency bins. Slope values are steeper at less distorted volume levels. As V_{Tx} approaches levels that overdrive and distort the audio, the slope becomes less steep.

The FSF method first calculates the periodogram of the audio received by the PTT device; the periodogram is a measure of the spectrum power. The periodogram is divided into 15 frequency bins of equal width and weight, with overlap between each bin, across a range from 200 Hz to 3250 Hz. This range is relevant to the majority of essential speech information, particularly after passing through a communication channel, which is defined as 300 Hz to 3400 Hz [5]. Dividing the periodogram into bins smooths the data and reduces the number of points that go through additional steps in the FSF process.

Next, the FSF algorithm measures the slope of the binned data by finding the bin with the highest power to use as a starting point. It is observed that the peak power occurs in the lower half of bins, 250 Hz to 1650 Hz, with the fundamental speech frequency falling in bins below the halfway point. The max is only searched for in the lower half of the frequency bins, which prevents errors if a recording is mostly noise and has a fairly flat periodogram. Designing FSF to use a max point in the lower half of bins was also important in order to prevent a case where there are not enough points for a linear fit. If a message is lost or is of suboptimal quality, the measurement could return unexpected results. Data collected with incomplete calls proved the safeguard to be effective, as low FSF scores were calculated. A linear fit of the upper, higher frequency bins is then performed; only the slope of this line is used in further calculations. The steepness reflects the impact that clipping due to an overdriven V_{Tx} has on the upper bins. Overall, this process provides a measurement of power in lower frequencies compared to higher frequencies and how that relationship changes after the audio is processed through the SUT. An overview of the FSF process is conveyed in Fig. 1. In short, FSF is defined as the ratio of the slope of the received audio's power across bins to the slope of the transmitted audio's power across bins.

Mathematically, the FSF process can be described as follows. Given periodogram signal power $p = [p_1, ..., p_n]$, p is divided into bins $b_k = [p_{L_k}, p_{U_k}]$, where $k = 0, ..., n$. Periodogram signal power p is used to calculate the average power per frequency bin, $\bar{p}_k = \frac{1}{U_k - L_k} \sum_{i=L_k}^{U_k} p_i$. Then the maximum $\bar{p}_j$ is found such that $\bar{p}_j \geq \bar{p}_k$ for all $k = 0, ..., n$. Using the data points $[\bar{p}_j, ..., \bar{p}_n]$, a line of best fit is given with slope m and y-intercept y_0. We define FSF in Eq. (1), where m_{Rx} and m_{Tx} represent slopes for received and transmitted audio data, respectively.

$$\mathrm{FSF} \doteq \frac{m_{\mathrm{Rx}}}{m_{\mathrm{Tx}}} \tag{1}$$

This ensures that the FSF of a signal with itself will always be one and provides a useful reference point for FSF values. This further focuses FSF values to represent the behavior of the SUT rather than the test signal. While any audio that travels through a communications system will not be a perfect replica of the original, it is expected that behavior that is similar to the original will have a similar slope value. It is worth noting that narrow band communication systems effectively contain a low pass filter, which rolls off high frequencies and can result in FSF values that are slightly greater than one for a channel without clipping. Furthermore, if the system has a steep, negative slope value and some clipping, it is plausible that there is a point where FSF $= 1$, but there is a slight amount of distortion. Because there is a fairly wide, flat "good" region that one would want to operate in, the algorithm looks for a plateau, not FSF $= 1$.

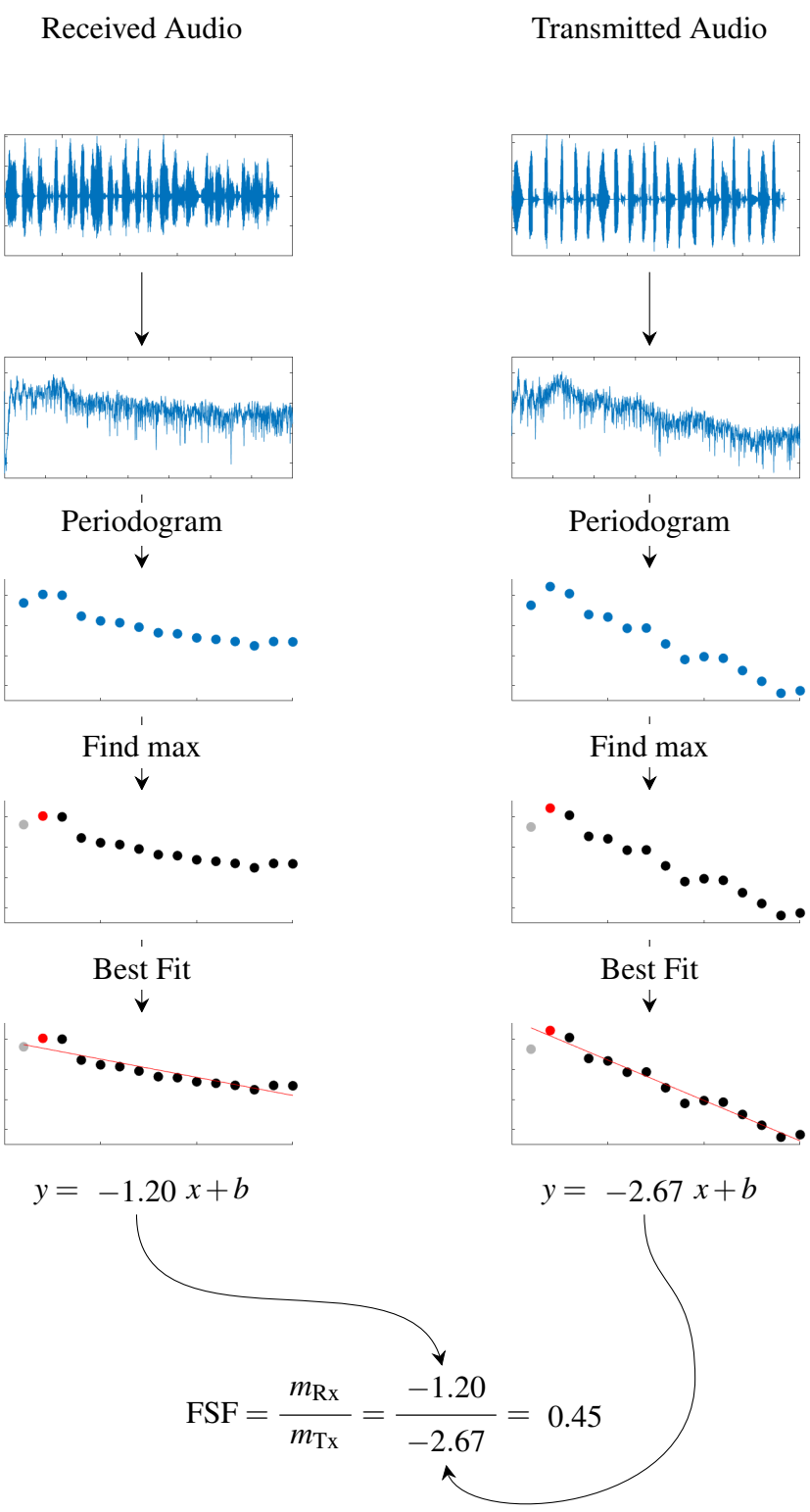


Fig. 1. FSF process. The received audio is recorded. The periodogram is calculated and divided into bins. The bin with maximum power is identified, and a best fit line to the right of that point is created. The slope of each best fit line is calculated. The y intercept is denoted with b, which is not used in FSF calculations.

3.2 Optimal Volume Plateau Identification

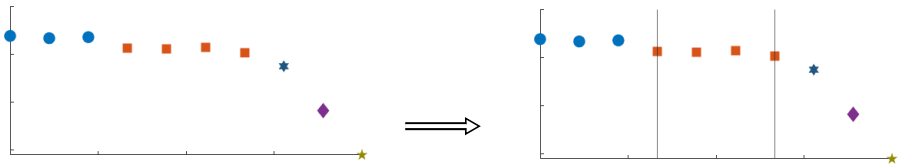
The goal of OVPIA is to identify the range of optimal volume levels that exist within the measured FSF scores as well as to select evaluation levels to drive the TVO. OVPIA was developed to look for the plateau of minimally distorted settings instead of a single location of peak quality. Ultimately, this plateau leads to an optimal V_{Tx}. As stated in the initial assumptions for measurement design, commercial MCPTT devices will not have a single optimal setting. Finding the edges of this range, the boundaries for optimal V_{Tx} options, will lead to a region of levels that are equally successful in performing QoE measurements. A V_{Tx} between these boundaries will be considered the optimal V_{Tx}. Across the range of evaluated V_{Tx}, the plateau of consistent FSF scores is evident. Outside this region, the FSF values either changed or formed smaller groups of data, while scores were lower outside the main plateau of stable values.

The variation of FSF scores across trials at the same volume is significant, as further described in Sec. 4.3. The approximate permutation test is used to determine statistically equivalent groupings of FSF data. A permutation test is a type of statistical hypothesis test designed to test whether or not two data sets are taken from the same distribution and is agnostic to the underlying distribution of the data [6]. It is performed by pooling and randomly relabeling the data and analyzing the new distributions a sufficiently large number of times, usually with $N \approx 1000$ resamples, to ensure adequate sample space coverage. The test can be used to classify if two distributions of data are equivalent and can take into account the natural variation in different measures. The OVPIA method uses the approximate permutation test to find groups of similar points. These groups are compared to find a large group of points with fairly high FSF scores.

The test starts out with no groups formed; the test interval, defaulting to $[-40,\ 0]$ dB, is sampled with the given number of initial V_{Tx} levels. Sets of FSF scores from a range of V_{Tx} levels are compared to determine if they belong to the same group. This way, noisy data can be accommodated to find a region of points that are "about" the same value. It is worth noting that "about" here means is determined by the variation in input (FSF) values. In order to detect the plateau, a certain number of initial volume levels are evaluated, defaulting to ten. The default number of initial volume levels was determined by running simulations with a variety of potential initial volume levels. Ten was selected as it balanced minimizing the number of evaluation levels needed to arrive at the solution without having a significant effect on final output variability.

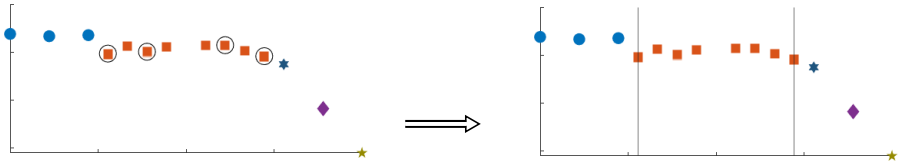
The grid spacing is set to an equal distance between sample levels. Once the initial levels have been evaluated in the TVO, groups are formed. The sample levels are evaluated one by one; each level's data is checked against the existing groups to see if it matches using the approximate permutation test. If the level matches, it is added to the matching group; if not, then it is placed in its own group. To further refine the search window, the grid spacing is halved so new sample levels are between the previous levels. If no groups have multiple levels, then the whole interval is re-sampled with the new grid spacing. Otherwise, the "best" group is chosen and the endpoints of the group are used for the new interval. The new sample levels are located one new grid spacing on either side of the new interval bounds.

The audio interfaces used in the SUT allow a user to adjust the V_{Tx} in integer increments, as such 1 dB was selected for the default tolerance. Additional V_{Tx} are evaluated until the tolerance threshold is met; when levels are less than 1 dB apart, the new level selection stops. Once the final "best" group is identified, the optimal level, the V_{Tx} a user should use for testing, is selected. The final "best" group becomes the optimal interval, and a single value is selected at the 80 % point in the interval. The process of selecting the 0.8 weight is described in Sec. 5. Figure 2 provides an overview of the grouping and level evaluation process.



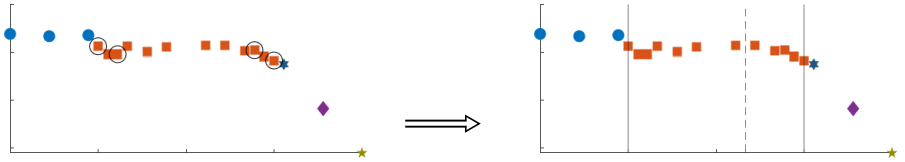
Take initial 10 points. Break into groups.

Pick the "best" group and sample at the interval boundaries.



Take an additional 4 points. Group new points.

Pick the "best" group and sample at the interval boundaries.



Take an additional 4 points. Group new points.

Pick the "best" group. Find optimal point from interval boundaries.

Fig. 2. OVPIA group selection process. The initial volume levels are evaluated. Throughout the process, more levels are evaluated, and average FSF points are added to groups until the optimal interval is found.

While the data presented in this publication typically demonstrates an optimal range of values with FSF scores closer to one, this is not always the case. It is worth noting that while the FSF score of the originally transmitted test audio would be one, this does not imply that the measurement system is looking for points where FSF $= 1$. It is also worth

noting that the data is unique to the SUT and thus groups may behave differently than the example cases in this publication.

4. Evaluating Developed Measurements

It is essential to ensure that the newly developed measurement methods deliver results that are repeatable and accurate. The TVO is used as the first calibration step to perform MCV QoE measurements. Once FSF and OVPIA were developed, they needed to be integrated into the overall TVO measurement system. Their effectiveness separately needed to be evaluated in order to ensure that they functioned in the desired use case.

4.1 Audio Clip Design

Audio clips were designed to test the range of distortion that could be caused by an overdriven V_{Tx}. Steps were taken to design audio files that would be sensitive to V_{Tx} and help identify markers of distortion caused by this parameter. Each audio file contains the top twenty loud words within the Modified Rhyme Test (MRT) database [7], by talker. A check was performed for good neighbors; a good neighboring word does not come from the same MRT keyword batch as the previous or next word in the audio file. These audio files utilize the same variably spaced speech technique used in Ref. [3]. With neighboring word checks and silence spacing, MRTs may be performed in post-processing on data if a user is interested in the impact of V_{Tx} on intelligibility.

Loud keywords were selected in order to test the worst case scenario for clipping if normalization was not utilized. Four talkers were used, two male and two female. Loudness was calculated using the A-weighting filtered value of each word within the database. The selected keywords are passed through an A-weighting filter and normalized to the mean A-weighted value of all MRT keywords, -34.3 dBA. In this context, dBA refers to the power of the audio after it is passed through an A-weighting filter. The mean of all keywords was selected in order to have a reasonably safe value that was unlikely to clip in tests with mid-test range volume levels. All four audio files follow the same structure.

4.2 Data Collection

Prior to developing the updated TVO, preliminary work was done to establish best practices to better understand the problem. Data was collected using the existing M2E latency measurement system [1] to obtain audio with a variety of V_{Tx}, allowing for research of methods to capture the impact of overdriven speech on received audio recordings. After using the recordings from the M2E latency data to design FSF and OVPIA, additional data was collected using the TVO script to verify success in the desired use case. All data collected in this project used the same radio model and audio interface settings aside from the parameter under test, V_{Tx}. In order to be robust to a variety of PTT technologies, data was collected for analog direct, Project 25 (P25) direct, and P25 trunked Phase 2 modes. The test setup is typical to MCV measurement systems, with the only change being the

use of a 1:1 transformer on the transmit radio path, as shown in Fig. 3. Previously, a 12:1 transformer was used, however a 1:1 transformer was utilized in order to allow evaluation of higher volume levels. Additional information on the test setup and components can be found in the access time paper [2].

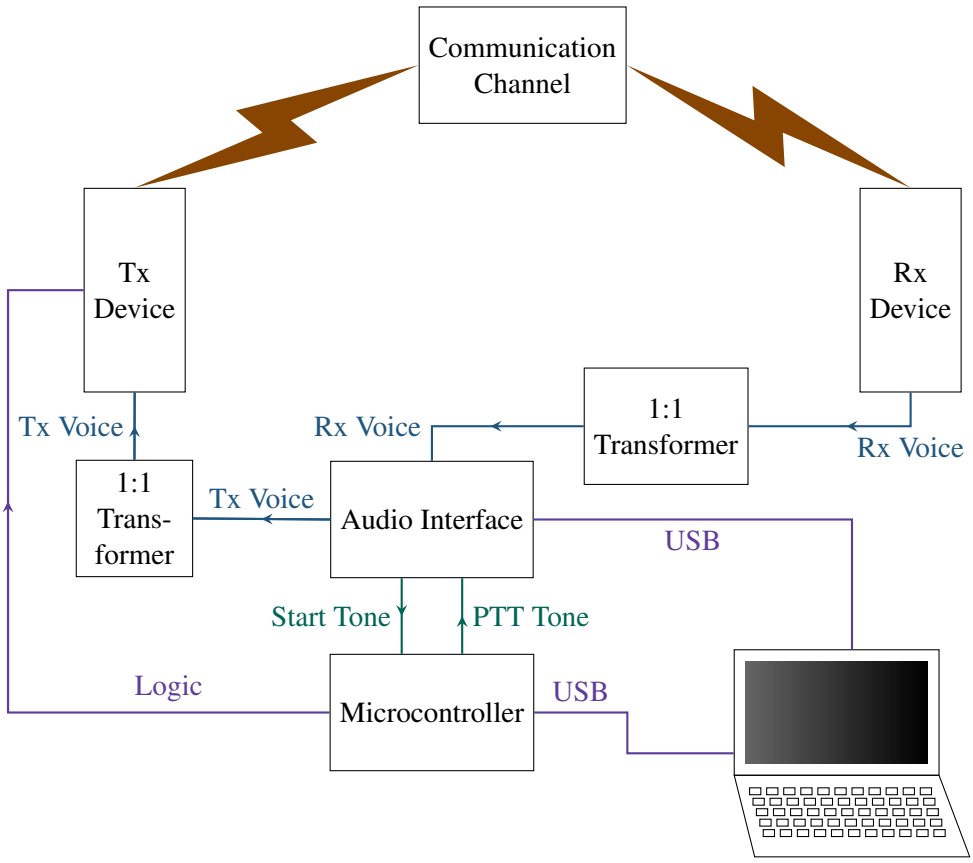


Fig. 3. Diagram of the measurement system setup

Final data was collected using the TVO. The TVO allows for the ability to automatically scale the waveform to V_{Tx} as selected by the algorithm. One can use TVO to take measurements at fixed, user-selected levels, as well as letting the test pause so that the user can manually adjust to the requested V_{Tx}. For the evaluation of the newly developed algorithms, perceptual evaluation of speech quality (PESQ) was measured in test data alongside FSF. The use of PESQ as a verification metric is described in Sec. 6.2. The output of the TVO displays a plot of average FSF scores across V_{Tx}, as in Fig. 4, as well as the optimal V_{Tx}.

Prior to any processing, a few observations were made while collecting data. Differences in the received audio recordings could be heard when testing P25 systems compared to analog direct. When higher V_{Tx} levels were used, P25 recordings sounded significantly louder for the first two keywords then leveled out as the trial progressed. This behavior was not witnessed on analog direct recordings, thus highlighting differences between

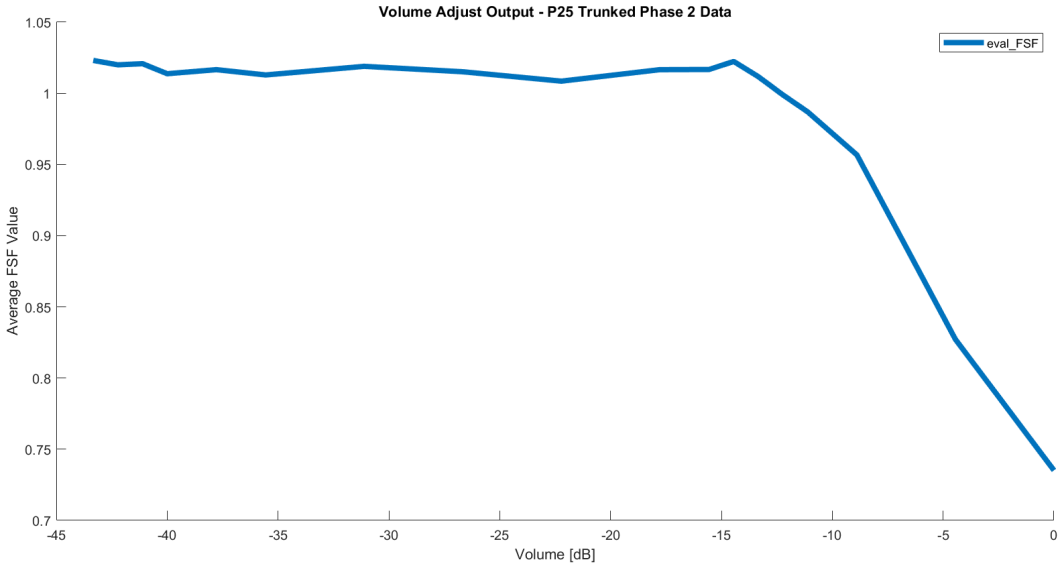


Fig. 4. Example output of the TVO. Average FSF values are plotted vs transmit volume levels.

technologies. Additionally, this behavior difference provides an example of PTT devices compensating for impairments.

4.3 Evaluating Frequency Slope Fit

The algorithm was used in numerous simulations to evaluate the behavior of FSF within the measurement system. These simulations allowed for making modifications during the improvement process, simulating the changes instead of going through lengthy data collection. Such simulations were especially useful for evaluations performed to fine-tune measurement method variables, such as band weights, the fitting of band values, and the type of frequency binning. Once finalized and integrated into the TVO, data was collected, as described in Sec. 4.2, and further evaluated using the same simulation tools as well as statistical analysis.

It is essential that FSF scores, the measurement that guides other parts of the TVO, are fairly consistent. The variance of scores across technologies was calculated to ensure that scores are not vastly different at an equivalent V_{Tx}. A contributor to this variation is that four talkers are used for the test audio clips. Figure 5 demonstrates an example of score clustering by talker, particularly evident in analog direct data. This clustering is a realistic scenario, as PTT users will speak with a variety of frequencies and annunciations of words. Where one talker begins to see a decline in quality caused by increased distortion levels at a certain V_{Tx}, another may have minimal distortion. Figure 6 shows the variance of average FSF scores over the range of evaluated intervals across all technologies. It is worth noting that the variance increases as clipping begins. Note that for all technologies evaluated in the SUT, the variance in the collected FSF scores is less than 0.013, as seen in Fig. 6; this value indicates stability of the measurement. This stability is true for all technologies

examined in the SUT including analog direct, even though it is the technology with the greatest variance.

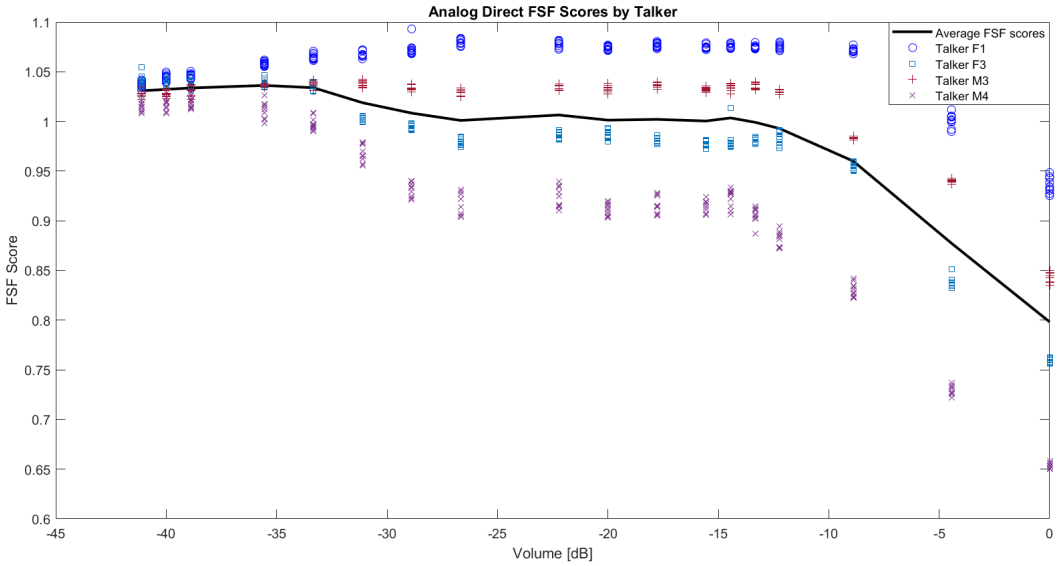


Fig. 5. Example FSF scores by talker. In analog direct data, the variation of FSF scores by talker is more evident than P25 technology data.

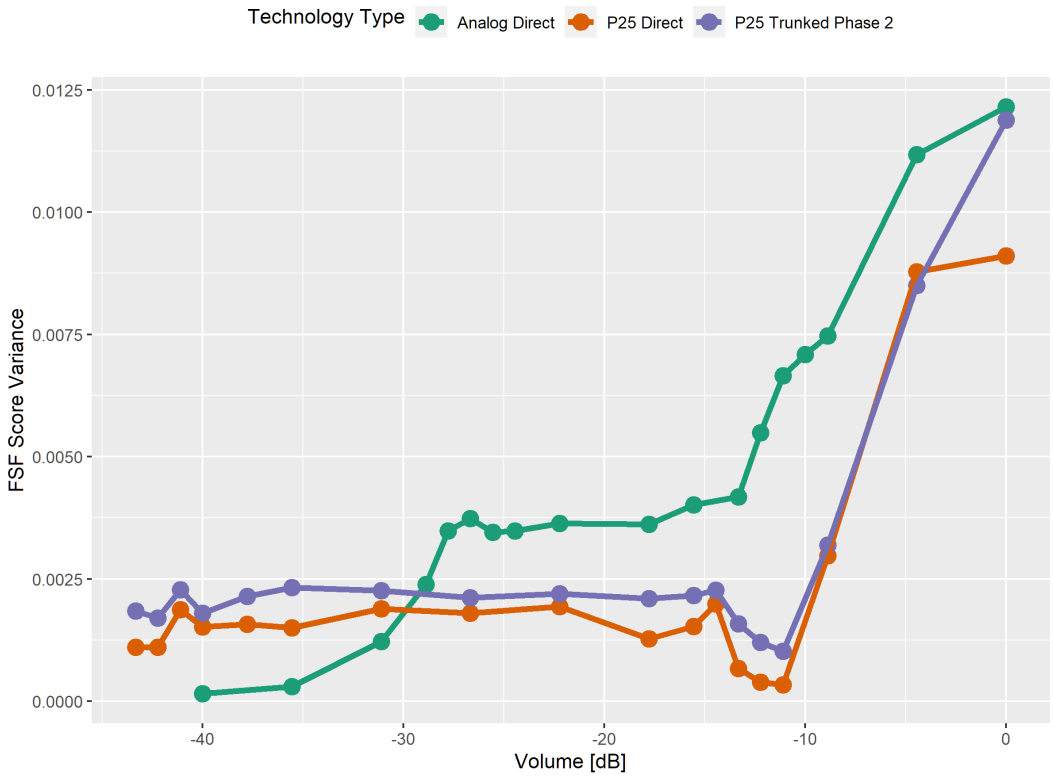


Fig. 6. Variance in FSF scores for each technology across transmit volume levels.

4.4 Evaluating Optimal Volume Plateau Identification

Testing the plateau detection method, OVPIA, presented a challenge because the algorithm chooses the points to evaluate and these are not known ahead of time. To simulate something close to what would be expected in an actual test, curve fits of FSF and PESQ values were created. To simulate the variation in values, noise with a given standard deviation was added to the FSF scores. This noise model does not reflect what is seen in reality, as the standard deviation will fluctuate as distortion varies across V_{Tx}.

Simulations were run using test data to evaluate how the algorithm handled a range of system noise. Tests were run adding noise with a standard deviation from 0.0075 through 0.125. From this extensive testing, it was found that the algorithm struggled with unrealistically low levels of noise. This behavior occurs because as the standard deviation of the noise goes to zero, so does the difference between the two means that the approximate permutation test considers to be the same. To fix this issue, dithering was added to provide a small amount of noise to the incoming data. While testing the effectiveness of OVPIA, it was noted that the results from very low noise test cases in simulations had higher standard deviation values for the optimal interval boundaries. As a safeguard, the output FSF values are dithered within OVPIA to handle a SUT with low noise. Based on the results of simulating a variety of noise levels, the dither value was set to 0.05 on the FSF scale. The dither

value has the effect of putting a minimum bound on how close values can be and still be treated as different. This improved accuracy and reduced the standard deviation of the final optimal interval values, with no negative impacts found. Dithering especially improved the stability of the upper end of the interval, the part of the range especially susceptible to variation due to it being a point of transition as volume levels begin to cause clipping.

By default, the TVO script evaluates levels between the range [-40 dB, 0 dB], as well as points adjacent to this range. The lower bound for P25 technologies within the SUT goes below -40 dB, which is possible because as the algorithm incorporates nearby points, these points may be added to the final optimal group. Also, note that FSF was designed to measure the amount of distortion in the communications system due to overdriven audio caused by the V_{Tx}, and it is expected that volumes on the lower end of the test input range contain minimal amounts of this distortion type. In some tests, this makes locating a lower endpoint to the interval difficult, as demonstrated by the slightly higher uncertainty values for the lower bound shown in Table 1. Overall, the stability of both bounds combined with minimal variance of FSF scores over selected intervals across all technologies indicates that the TVO is stable.

Table 1. Mean and expanded uncertainty of upper and lower interval boundaries. Calculated using 10 trials and coverage factor of k = 2.26 for each P25 technology and 30 trials and a coverage factor of k = 2.05 for analog direct.

Technology	Lower Bound [dB]	Upper Bound [dB]
Analog Direct	-33.5 $\pm$ 2.3	-11.2 $\pm$ 1.0
P25 Direct	-43.3 $\pm$ 0.7	-12.1 $\pm$ 1.0
P25 Trunked Phase 2	-43.0 $\pm$ 0.8	-12.6 $\pm$ 1.3

5. Optimal Volume Level Identification

The two major components of TVO are the measurement method, FSF, and the optimal point selection algorithm, OVPIA. For the original version of TVO described in Ref. [1], a golden section search was used to find the optimal V_{Tx}. Golden section searches expect one local maximum on the interval, which would sometimes cause the search to get hung up on system variation and focus on V_{Tx} values near each other instead of looking at a wider range. Realistically, there is not just one optimal level where device settings work well. It aligns with a practical device characterization to have a range of levels to pick from. The new algorithm checks across a wide range of V_{Tx} to capture this optimal range and where behavior begins to change as V_{Tx} approaches levels that are overdriven.

When determining the best way to calculate the optimal V_{Tx}, more emphasis was placed on the upper interval bound. Realistically, one would want the signal to be as strong as possible with minimal distortion while performing measurements. While talking into a PTT device, a user would typically not whisper; the case presented here focuses on a V_{Tx} between an average talking voice and yelling to the point of distortion. Additionally, the

distortion levels rise sharply in the upper V_{Tx} range where clipping occurs, creating an edge that is more obvious for the algorithm to detect. Inherently, as FSF measures overdriven audio caused by V_{Tx}, scores decrease sharply in the upper V_{Tx} range as clipping begins. This transition point, and a steeper slope, creates an edge that the algorithm recognizes as not part of the optimal region. To this end, a weighting process was utilized, with the upper bound weighted more heavily than the lower bound.

To rigorously evaluate and optimize this weight, an algorithm was applied during statistical analysis to search through different weights, ranging [0.5, 1], and determine the optimal V_{Tx} based on FSF scores. Then, the algorithm verified that the interpolated PESQ at the optimal V_{Tx} was statistically equivalent to the maximum observed PESQ score within each trial. Statistical equivalence is determined by whether or not the confidence interval of the difference in PESQ estimates and maximum observed PESQ contains zero. If all of these confidence intervals contain zero, then the optimal weight is one, meaning that the upper bound should be used. If some do not contain zero, then the optimal weight will be less than one. More occurrences of confidence intervals not containing zero result in a lower optimal weight value. This method verifies that the optimal V_{Tx} provides sufficient PESQ scores.

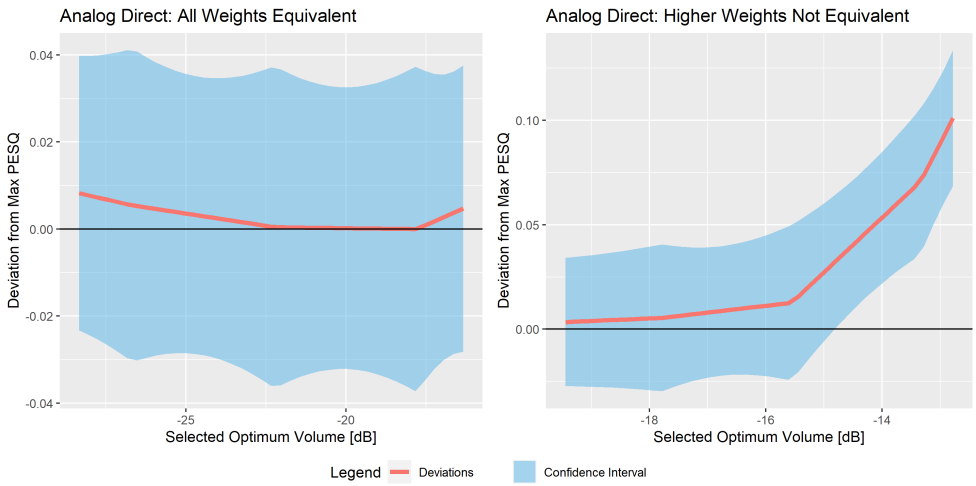


Fig. 7. Example optimal volume weights. Optimal volumes for two different trials of analog direct technology across different weightings. The left edge corresponds to weight 0.5 and the right edge weight 1. The black line at zero deviation is fully contained in the confidence interval on the left, but not on the right.

To better illustrate the mathematical process for determining the optimal weight for each technology, consider the following. Let the difference in audio quality be represented by $D_j = P_{\max} - P_j$, where P_j is the PESQ score at j^{th} tested volume level, V_j, and $P_{\max} = \max(P_1, P_2, ..., P_{N_v})$, where N_v is the number of volume levels tested for a technology. The V_j are all between the lower and upper volume interval bounds, V_L and V_U, determined by OVPIA. Each D_j has an associated V_j, representing the volume at which the

data was taken. If the 95 % confidence interval for a D_j contains zero, then it is statistically equivalent to zero. For each trial, t, in a technology, the maximum D_j statistically equivalent to zero is found, represented by $D_{\max,t} = \max(D_j | D_j \text{ equiv to } 0)$. It follows that for each $D_{\max,t}$, there is an associated best volume, V_t. The relative location of V_t within its plateau is calculated by scaling the OVPIA-generated lower and upper volume bounds to 0.5 and 1, respectively, and is stored as the best weight, w_t, for that trial. Finally, the best average weighting value for a technology, w_{tech}, is calculated by averaging the best weight, w_t, across all trials per specific technology,

$$w_{\text{tech}} = \frac{1}{N_T} \sum_{t=1}^{N_T} w_t,$$

where N_T is the number of trials for a technology. This establishes the best average weighting value for a technology. The average of the weighting values, w_t, was chosen as the optimal technology weighting because the weighting values were tightly clustered around the mean value. Due to this clustering, the minimum of the weighting values is close to the mean value. In order to ensure that the selected weight will work for every technology, it was decided that after the weighting process was implemented for each technology, a conservative value lower than the observed minimum would be selected and applied to all technologies.

Table 2. Optimum weights for each technology. Determined by averaging optimum weights across trials for each technology within the SUT.

Technology	Optimum Weight
Analog Direct	0.83
P25 Direct	0.89
P2 Trunked Phase 2	1.00

The weighting algorithm was repeated across the entire final data set, producing the results in Table 2. Across all P25 trunked Phase 2 data for the SUT, the optimum weight was determined to be one. For P25 direct and analog direct technologies, at higher V_{Tx} some of the confidence intervals of the differences in PESQ estimates and maximum observed PESQ did not contain zero, as shown in Fig. 7. This effect reduced the average optimal weights for those technologies within the SUT to 0.885 and 0.826, respectively. Attentive to this varied performance, a conservative weight of 0.8 was selected. This value ensures that the optimal V_{Tx} provides a strong signal that is still safely within the range of minimal distortion that will safeguard for any technology evaluated using the MCV QoE measurement system.

6. Transmit Volume Optimization Finalization and Evaluation

The purpose of the TVO is to identify an optimal V_{Tx} within a MCV QoE measurement system. The values described in this publication are for verification purposes and reflect

the NIST PSCR Mission Critical Voice team's laboratory environment and SUT. While individual scores and output for a given SUT may vary, the process will remain. The updated TVO will deliver a single V_{Tx}, identified from a range of values where settings are not clipping. By running the updated TVO, the impact of V_{Tx} on measurement uncertainty for other QoE measurements will be minimized.

A variety of verification metrics were applied to validate the updated TVO. Since the output will impact the results of all MCV measurements, it is essential that the updated TVO consistently gives an optimal volume that results in high quality audio. Its stability was evaluated by examining the optimal V_{Tx} across tests, as well as across the intervals selected. To validate that an accurate optimal volume level was given, PESQ was used to validate that volume levels with high quality audio were found.

6.1 Transmit Volume Optimization Stability and Uncertainty Analysis

In the uncertainty analysis, repeated tests were performed on each technology and the collected data was used to evaluate optimal V_{Tx} values for technologies both individually and relative to each other. Therefore, 10 sets of TVO data were collected for P25 direct and P25 trunked Phase 2, while 30 sets were collected for analog direct because its measurement output had a higher variance. The stability of optimal transmit volume intervals identified by OVPIA was assessed using 95 % confidence intervals. In post-processing, a weighting process was applied to the data to calculate the optimal V_{Tx} within the interval. To ensure that high-quality audio was produced at those levels, audio quality was validated using the data's PESQ scores. As described in Sec. 5, the previous version of the TVO utilized a golden section search and audio quality measurement. While the new version of the TVO has a different approach than its predecessor, validating it using previous measurement methods was an important verification step.

Evaluating the stability of the package for each technology is done by characterizing the uncertainty of the measurement, starting with the mean of the optimal volume across the data. Repeated measurements were taken and uncertainty was calculated as shown in Table 3. Optimal transmit volume levels for each technology measured within the SUT have an associated uncertainty of less than 2 dB. The spread of optimal levels varies by technology, with analog direct having a wider distribution of optimal values, reflected in the uncertainty. Furthermore, it can be seen in Fig. 8 that the data is clustered. The analog direct data has two main clusters while most of the data for P25 trunked Phase 2 is clustered near -18 dB. P25 direct has three distinct clusters in close proximity of each other. These spread values indicate that the mean is an accurate representation of the data. Based on these findings, the TVO delivers a reasonably stable output. This stability is essential, as the settings provided are used throughout all other MCV QoE measurements.

Additionally, since each interval is contained by a lower and an upper bound, it is reasonable to say that stability of the bounds implies stability of the optimal point, which is provided by the algorithm to avoid the user arbitrarily choosing a point on the interval for testing. It is shown via the uncertainty analysis in Table 1 that the upper and lower bounds

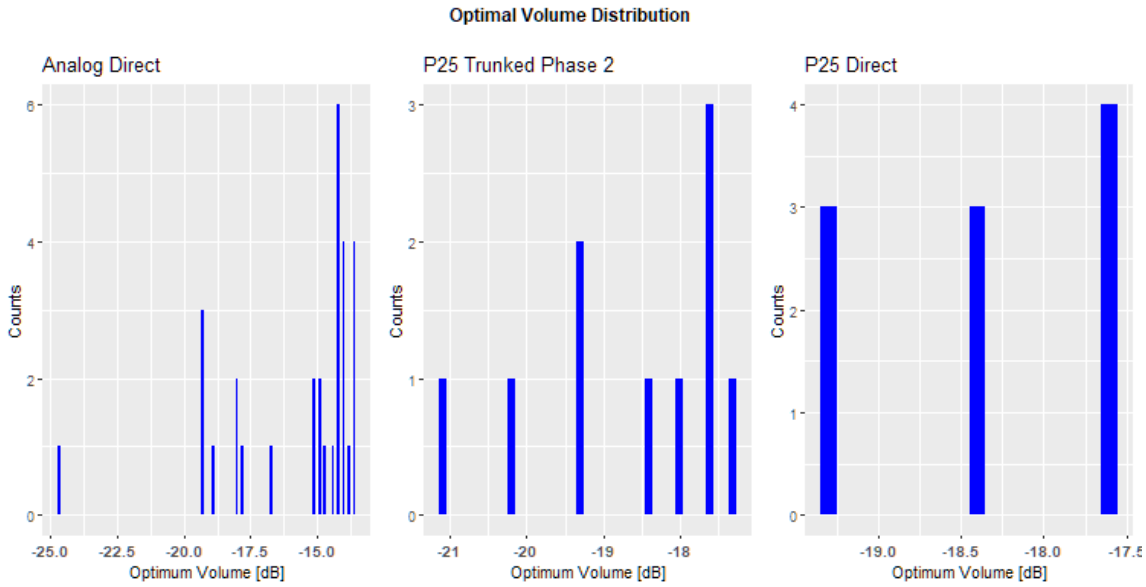


Fig. 8. Optimal volume distribution. Optimal volume distributions for the SUT using 0.8 weighting factor for analog direct (left), P25 trunked Phase 2 (center), and P25 direct (right).

Table 3. Mean and expanded uncertainty of optimal transmit volumes. Calculated using coverage factor of k = 2.26 for each P25 technology and a coverage factor of k = 2.05 for analog direct.

Technology	Optimal Transmit Volume [dB]
Analog Direct	-15.7 $\pm$ 1.0
P25 Direct	-18.4 $\pm$ 0.6
P25 Trunked Phase 2	-18.6 $\pm$ 0.9

for each technology have low uncertainties and thus are stable. For the upper boundary of the interval, the expanded uncertainty is on the order of 1 dB. The upper interval is of most interest from the standpoint of optimal V_{Tx} selection, as the highest possible volume that avoids distortion is ideal.

6.2 Audio Quality Verification

It is important to verify that high quality audio is produced at the optimal V_{Tx}. OVPIA selects V_{Tx} to evaluate based on the process described in Sec. 3.2. Thus, it is likely that audio was not recorded at the suggested optimal V_{Tx}, as the optimal is based on the interval and not the individually evaluated levels. To address this scenario, a linear interpolation between the two nearest evaluated levels was used to obtain an estimated PESQ score for the optimal V_{Tx} level. Demonstrated in Fig. 9 is a case where the optimal V_{Tx} is not an evaluated point, and thus the two nearest points were used for the linear interpolation process.

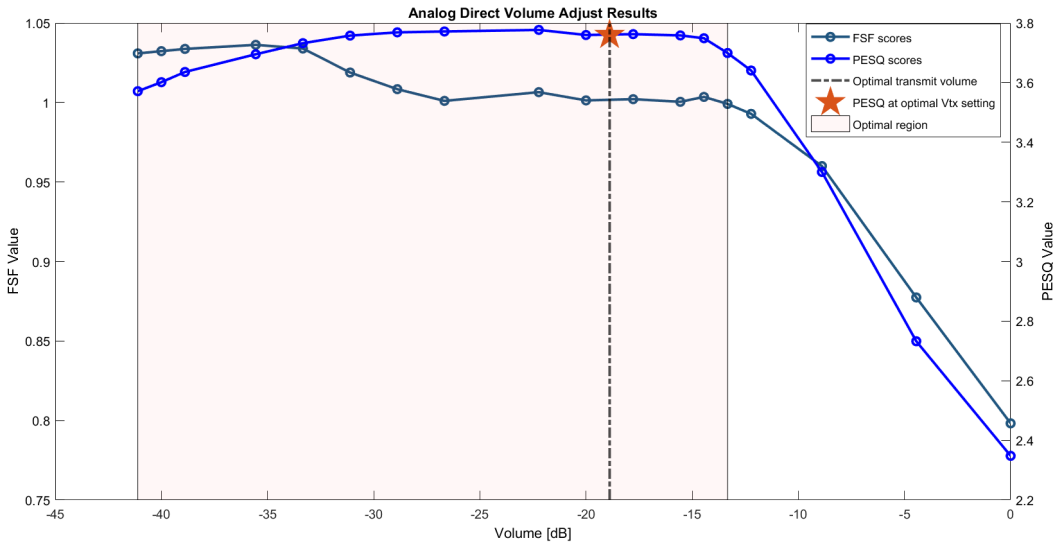


Fig. 9. Example analog direct TVO data. The evaluated FSF and PESQ scores are plotted. The optimal transmit volume for this data set lies between two evaluated points. The PESQ value used for verification is determined using linear interpolation.

Once obtained, the PESQ estimate was compared to the maximum observed PESQ score on the plateau of each data set. For strict analysis, a 95 % confidence interval was calculated around a single value, the averaged difference between maximum observed PESQ and PESQ estimates, for each technology. This difference is calculated by first finding the difference between maximum observed quality and quality estimates for each trial, and then averaging those differences before developing the confidence interval. As exhibited in Table 4, this confidence interval does not contain zero. This range is important because a confidence interval containing zero indicates no statistical difference between quality estimates and maximum observed quality. Although the confidence interval does not contain zero, this result is expected because the difference is positive by definition and because the standard error is minute. Furthermore, the difference within the confidence interval is less than 1 % in the worst case scenario. Notice that a value close to zero, less than 0.04, is obtained for the lower bound of the 95 % confidence interval for each technology measured in the SUT in Table 4. On the PESQ scale, this value is reasonably small and the functional difference is not significant from the realistic measurement perspective.

To further verify that no functional difference exists, a secondary statistic was calculated as shown in Fig. 10. For each technology, it describes the difference in the mean of the estimated PESQ scores at optimal volume to the mean of the maximum observed PESQ scores across trials. The means of the two quantities are calculated first, then the confidence interval is built around the difference of those means. Although less strict, these confidence intervals contain zero, thus indicating no statistical difference between audio quality estimates at optimal volume and highest observed PESQ values across trials. This statistical equivalence confirms that the updated TVO consistently identifies a V_{Tx} with minimal dis-

tortion that also produce high quality audio.

Table 4. Average difference in PESQ scores between observed maximum and estimate at optimal volume settings. 95 % confidence intervals included in parentheses.

Technology	PESQ Difference
Analog Direct	0.037, (0.022, 0.052)
P25 Direct	0.026, (0.019, 0.032)
P25 Trunked Phase 2	0.015, (0.011, 0.018)

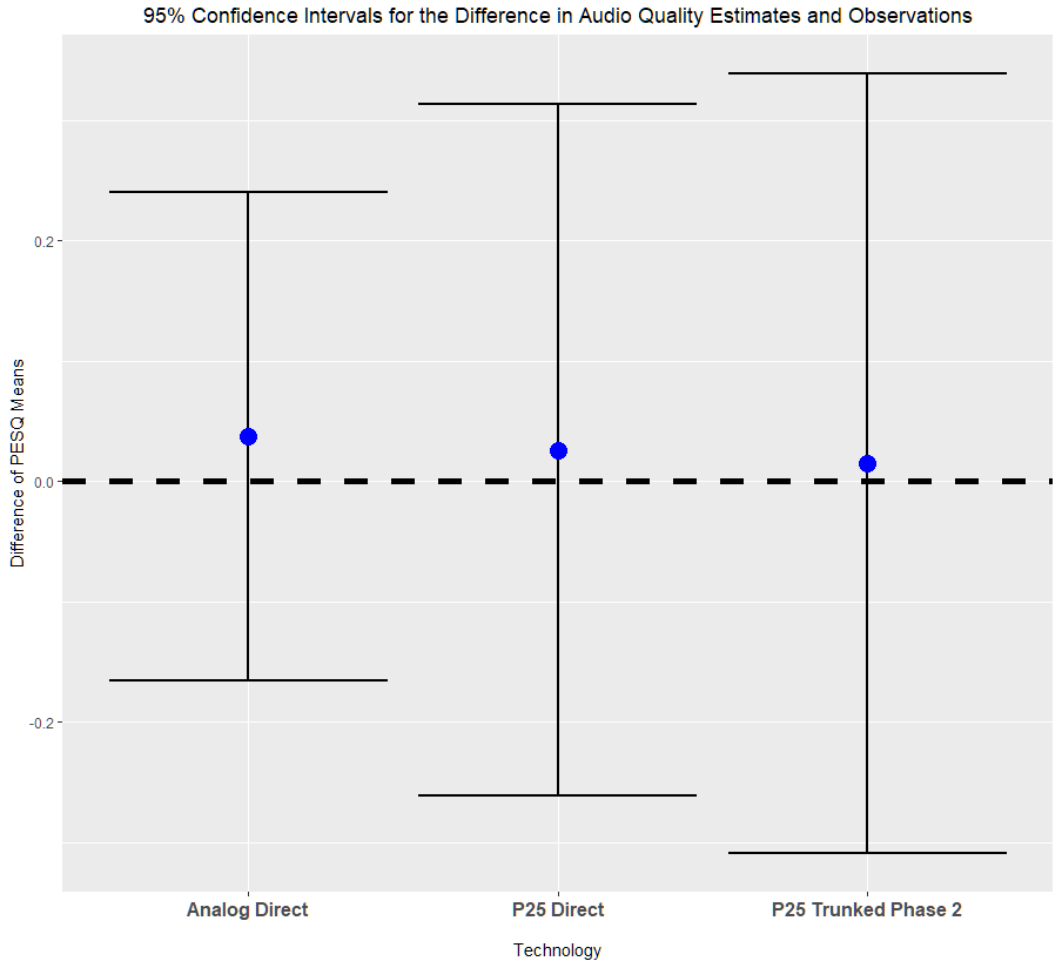


Fig. 10. Confidence interval analysis. Confidence intervals for the difference between mean PESQ observed and mean PESQ estimate at optimal volume. Each point represents the difference between observed and estimated values for that technology and is the center of the interval.

7. Conclusion

This project focused on developing a reliable, repeatable calibration step for MCV measurements by determining an optimal V_{Tx}. The updated TVO described in this publication provides users with a single, reliable V_{Tx} that minimizes uncertainty and improves quality for additional MCV QoE measurements. Optimal V_{Tx} for a PTT communications system can be determined by establishing the amount of distortion caused by overdriven transmit audio across a series of potential levels. By minimizing impairments caused by V_{Tx}, additional MCV measurements obtained will have less uncertainty contributed by this parameter. In order to improve the TVO, two main tools were developed. FSF characterizes clipping caused by overdriven speech through a transmit device. FSF compares the original transmitted audio to the received audio to measure how this distortion impact propagates through the communications system. OVPIA is a plateau detector that finds stable regions of V_{Tx} settings based on FSF scores. This stable interval is used to select the optimal V_{Tx} within a range of levels with minimal distortion. The MCV QoE measurement system user can run the TVO for their specific SUT and identify the optimal V_{Tx}.

Steps were taken to ensure the output of this package are repeatable and consistent. Individual results will vary by SUT; the stability analysis demonstrated in this paper validates stability of the TVO measurement. The optimal V_{Tx} level is selected from within a stable region of FSF values. This value, with a 0.8 weight on the upper end of the interval, was selected based on statistical analysis. This value ensures that the optimal V_{Tx} provides a strong signal that is still safely within the range of minimal distortion. In order to ensure that high audio quality is produced in the identified optimal interval, PESQ was used as a validation tool.

MCV researchers at NIST PSCR continue to minimize the cost of internally-developed measurement systems while improving the distribution of tools to a variety of users. In addition, MCV researchers will release a Python version of the TVO and verify functionality with PTT devices that use broadband.

References

[1] Frey J, Pieper J, Thompson T (2018) Mission Critical Voice QoE Mouth-to-Ear Latency Measurement Methods (NIST), IR-8206. doi: 10.6028/NIST.IR.8206

[2] Pieper J, Frey J, Greene C, Soetan Z, Thompson T, Voran S, Bradshaw D (2019) Mission Critical Voice QoE Access Time Measurement Methods (NIST), IR-8275. doi: 10.6028/NIST.IR.8275

[3] Greene C, Frey J, Soetan Z, Pieper J, Thompson T (2020) Mission Critical Voice QoE Access Time Measurement Method Addendum (NIST), IR-8328. doi: 10.6028/NIST.IR.8328

[4] Bucci D (2017) *Analog Electronics for Measuring Systems* (ISTE Ltd and John Wiley and Sons, Inc.), 1st Ed.

[5] Alliance for Telecommunications Industry Solutions (2019) Atis telecom glos-

sary - voice frequency (vf). URL https://glossary.atis.org/glossary/voice-frequency-vf/?char=V&page_number=12&sort=ASC.

[6] Chung E, Romano JP (2013) Exact and asymptotically robust permutation tests. *The Annals of Statistics* 41(2):484–507. doi: 10.1214/13-AOS1090. URL https://doi.org/10.1214/13-AOS1090

[7] (2016) Modified rhyme test audio library, NIST. URL https://www.nist.gov/ctl/pscr/modified-rhyme-test-audio-library.

Appendix

A. Measurement System Implementation

The following subsections contain an overview of the testing process. Individual measurement setups may vary.

A.1 Setup

Download all necessary items, such as code and audio files.
Code and audio files are available at: https://github.com/usnistgov/MCV-QOE-TVO.
Data is available at: https://doi.org/10.18434/mds2-2432.
Please see supplies list in Ref. [2] and system diagram in Fig. 3.

A.2 Transmit Volume Optimization Procedure

The following procedure should be followed per each combination of PTT device and audio file that will be utilized for other MCV measurements. It is best to set V_{Rx} to the same value for every QoE measurement.

1. Set V_{Tx} to the maximum volume level, 0 dB, on the audio interface (see Fig. 11). When setting the transmit volume, be sure to unlock the audio channels (click the chain link icon) so only channel 1 is set. Channel 2 is the start signal volume and lowering could cause the start signal to be missed.
2. To set V_{Rx}, adjust the gain knob on the audio interface (Fig. 13.1). Use the 1-location M2E latency `test.m` from https://github.com/usnistgov/mouth2ear to play the test audio.
3. While running a few trials, adjust V_{Rx} to a level where no clipping is observed on the audio interface. If audio is clipping, the light under "clip" near the gain knob will be red. Clipping may not occur for every word or talker; check through a few trials.
4. Enter `volume_adjust.m` input parameters. The default settings may be used. Example input: `volume_adjust`.
5. Enter test start notes (Fig. 12). Be sure to note V_{Rx} used for future reference.
6. Run test. Use `volume_adjust.m` to identify an optimal V_{Tx}.
7. Listen to the output audio for a few trials to ensure the test is behaving as anticipated.
8. Upon ending, the test window will show a plot of the data as well as the optimal V_{Tx} (Fig. 12).
9. Set V_{Tx} to the suggested transmit volume in future testing. Be sure to record this value and the V_{Rx} used.

There are other audio interface controls to check to improve the test running experience. The PTT tone is also impacted by volume levels. The pad button (Fig. 13.6) must be on

and the gain knob (Fig. 13.5) should sit just below the level where the red clip light turns on.

The mix and main out knobs (Fig. 13.3 and Fig. 13.7) can be used to adjust what you hear from the speaker. Usually the mix knob is set to "IN" to play back recorded audio, this way it is easy to hear if audio is going through the devices. The main out volume controls the volume out of the speaker. A suitable setting is based on how loud one wants to hear test audio as well as the position of the volume on the speaker.

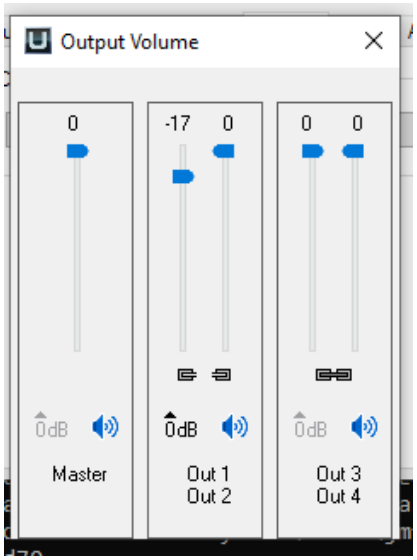
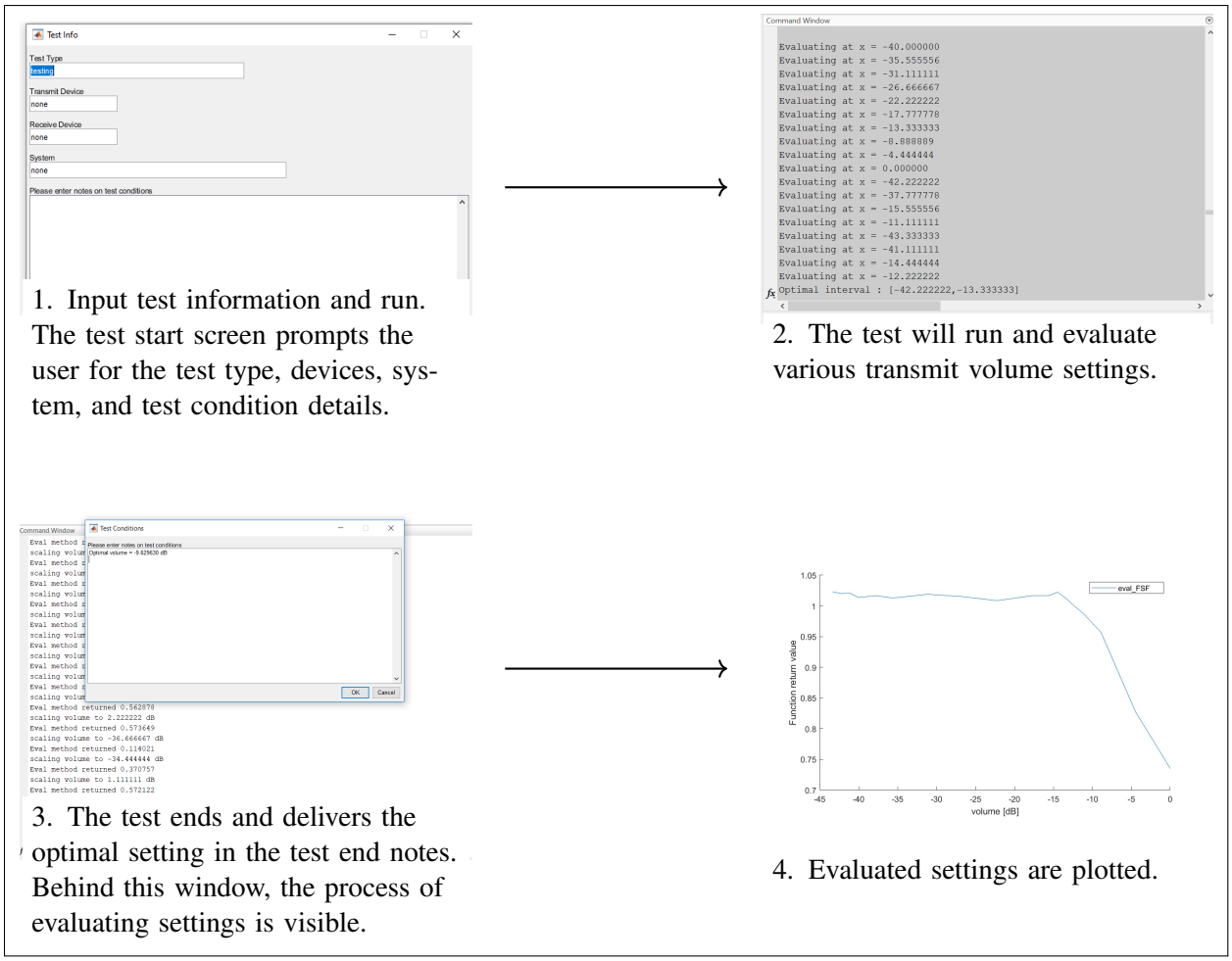


Fig. 11. Transmit volume controls. Channel 1 is of interest to setting the transmit volume.



1. Input test information and run. The test start screen prompts the user for the test type, devices, system, and test condition details.

2. The test will run and evaluate various transmit volume settings.

3. The test ends and delivers the optimal setting in the test end notes. Behind this window, the process of evaluating settings is visible.

4. Evaluated settings are plotted.

Fig. 12. TVO test process.

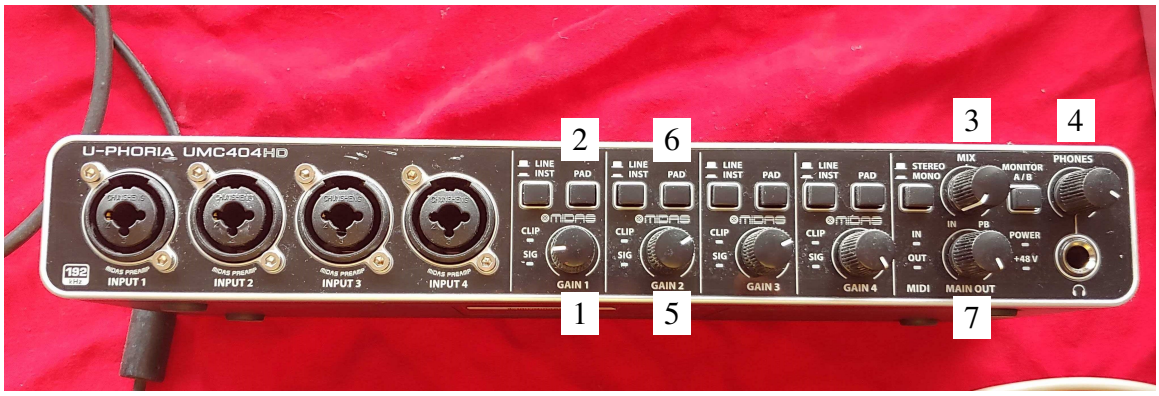


Fig. 13. Audio interface controls. The volume and audio control knobs on the front of the audio interface.
Labelled items include: **1.** Gain used to set V_{RX} **2.** Pad control for the RX device **3.** Mix knob **4.** Speaker control **5.** PTT tone gain **6.** Pad control for PTT tone **7.** Main out knob

B. Additional Figures

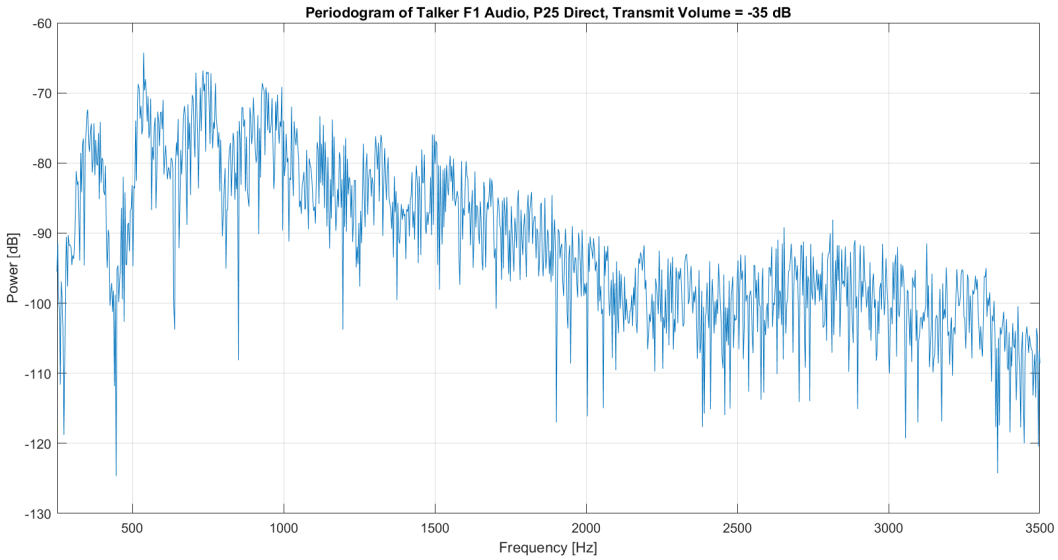


Fig. 14. Example periodogram using received talker F1 audio. Transmit volume set to -35 dB, P25 direct data.

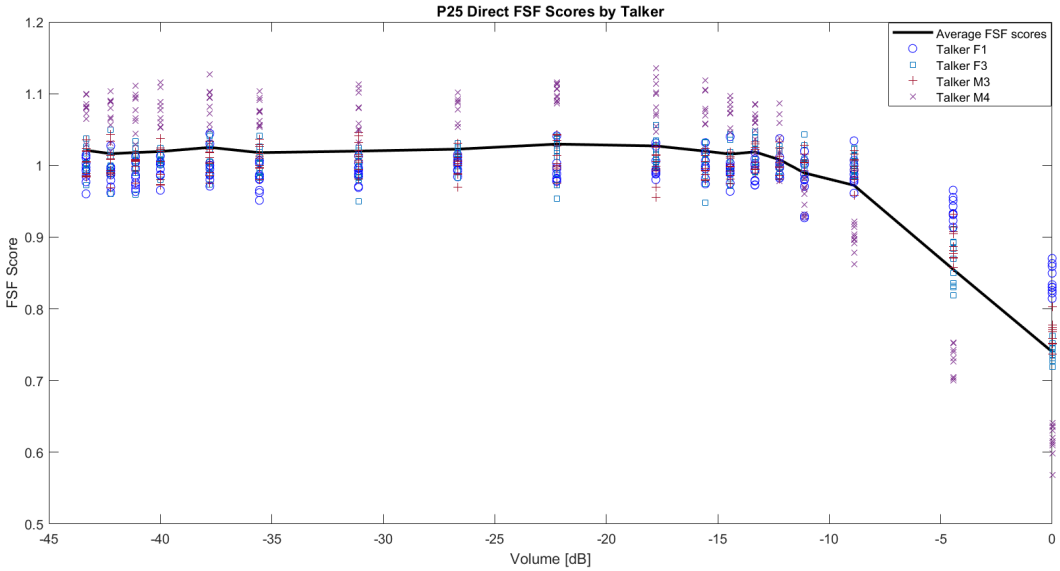


Fig. 15. FSF scores by talker for a set of P25 direct data.

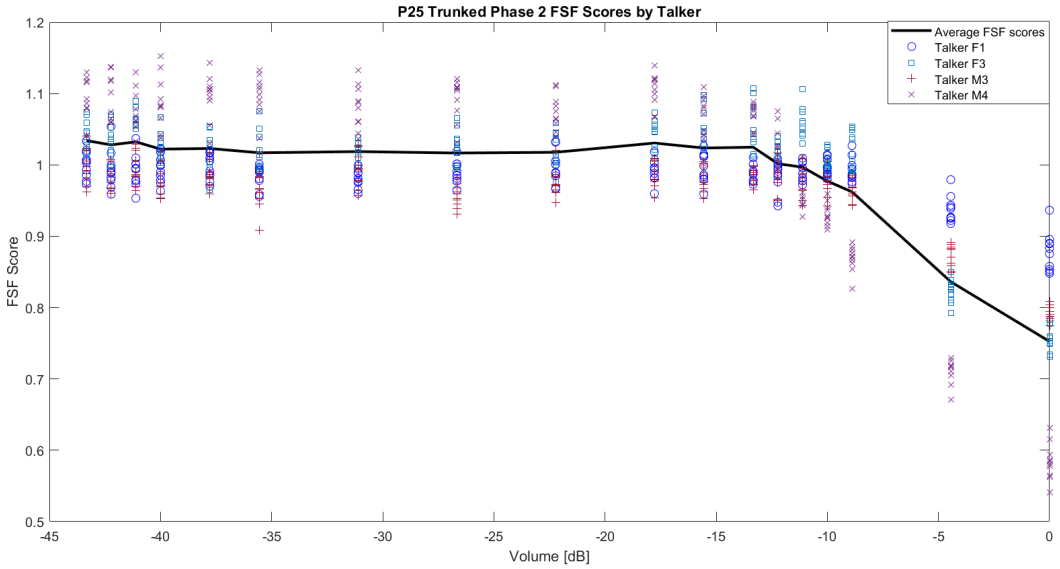


Fig. 16. FSF scores by talker for a set of P25 trunked Phase 2 data.