NTIA Technical Memorandum 24-571

Joint Analyses of No-Reference Speech Quality Estimation Tools and Conference Speech Recorded in Diverse Real-World Conditions

Jaden Pieper
Stephen D. Voran

Technical Memorandum

NTIA Technical Memorandum 24-571

Joint Analyses of No-Reference Speech Quality Estimation Tools and Conference Speech Recorded in Diverse Real-World Conditions

Jaden Pieper
Stephen D. Voran

U.S. DEPARTMENT OF COMMERCE

Alan Davidson
Assistant Secretary of Commerce for Communications and Information
National Telecommunications and Information Administration

July 2024

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

CONTENTS

FIGURES

TABLES

JOINT ANALYSES OF NO-REFERENCE SPEECH QUALITY ESTIMATION TOOLS AND CONFERENCE SPEECH RECORDED IN DIVERSE REAL-WORLD CONDITIONS

Jaden Pieper, Stephen D. Voran[1]

Recently, prerecorded audio and video presentations, as well as virtual meetings, have become a common component of professional life, due to health and environmental considerations. This places new responsibility on participants to generate audio that is of sufficiently high quality to effectively communicate. This memorandum provides analyses of real-world audio from a virtual component of a 2023 conference which encompasses a wide range of recording environments and conditions. We use both signal analyses and novel machine learning-based no-reference speech quality estimators and we evaluate their performance relative to each other. We utilized NISQA, WAWEnets, and TorchAudio-Squim, and found that while their scores show only modest agreement, we can use each to successfully identify low-quality speech. Finally we offer remediation steps for speech conferencing, to avoid many of the impairments observed in this work.

Keywords: conference speech, intelligibility, no-reference speech quality assessment, speech impairments, speech quality

1 INTRODUCTION

Remote participation in meetings, conferences, and workshops has become much more common in recent years. This has been driven by concerns for public health and environmental protection and will likely continue into the future at some level. Remote participation frequently includes audio and video streams contributed by individual participants located in a wide variety of environments and using conveniently available equipment. We have observed that the resulting audio signals routinely range from very nearly studio-quality to highly impaired and practically unusable. We have also observed that many of the common serious impairments could be significantly mitigated by fairly simple steps.

Recent years have also brought a dramatic increase in the development and availability of no-reference (NR) tools for evaluating speech signals. NR indicates the ability to assess a delivered speech signal without any access to a clean original version of that speech signal, whereas full-reference (FR) tools require an original version. Early NR tools relied on signal processing (SP) alone (e.g. [1]–[3]). Machine learning (ML) has fueled the recent resurgence of innovation resulting in tools that produce an estimate of a subjective speech quality score (e.g. [4]–[17]) and those that also estimate objective speech quality values (e.g. [18]–[24]). A conferencing speech challenge at Interspeech 2022 [25] specifically addressed NR "speech quality assessment in online conferencing applications."

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, Colorado 80305.

This paper exploits the recent progress and availability of NR tools in order to evaluate the speech in real-world recorded presentations produced by over 2500 different contributors in 2023. In Section 2 we describe in detail how we accessed the recordings. We then apply an ensemble of SP (Section 3) and ML (Section 4) tools and discuss the results they produce. In Section 5 we report that it is not difficult to reliably identify recordings that are seriously lacking in one or more respects. Fortunately, it appears that simple changes that cost little or nothing could lead to substantial improvements in many cases.

The development and evaluation of speech quality assessment tools is typically accomplished through the use of specifically constructed datasets. This allows for the careful control of impairment levels and types, but can limit the scope and realism. We are not aware of any previous study of NR speech assessment tools that uses completely uncontrolled, highly relevant, real-world recordings as we have done here. Thus this memorandum contributes a novel study on NR tools, a novel study on quality and impairment in speech recorded for a major conference, and suggestions for reducing common impairments to facilitate better conference communications.

2 ACQUISITION OF CONFERENCE SPEECH EXCERPTS

A major 2023 electrical engineering research and development conference included online access to presenter-created video presentations as part of the conference materials. These recorded video presentations were produced in advance by individual participants in their local environs using available equipment, giving the potential for a unique study of real-world locally-produced speech recordings. Datasets typically used in NR tool development and evaluation use clean speech from a single, controlled recording environment while noise, reverberation, and other impairments are simulated with software. Work in [25] moved somewhat beyond this paradigm. For example, subjects recorded scripted voicemail messages through real phone connections in [10] and spontaneous speech while walking through outdoor and indoor public spaces in [26]. In contrast, we analyze samples from the population of real-world conference-related recordings. These recordings present unscripted speech recorded in uncontrolled environments, as selected and used by conference participants, without any assumptions or artificial motivations.

We played example portions of the two-channel audio component of each recording. These example portions ranged from 14 to 50 seconds in length. The mean length was 36 seconds and is about 8% of length of a typical full presentation. To avoid inappropriate server or network stresses, we spaced the play requests consistent with that of a single user. It was not possible to implement the desired rich ensemble of audio analysis tools in real time, so we developed a solution using re-recording and temporary audio files. Specifically, the playing of each excerpt was captured in a temporary audio file. We then applied our ensemble of analysis tools to each file before it was deleted. The temporary audio files were created as shown in Figure 1. Note that the digital-to-analog (D/A) converter in the USB audio interface captures the playback signals a user would normally consume with speakers, headphones, or earbuds. The D/A conversion uses two channels, a sample rate of 48 kHz and a bit-depth of 24 bits/sample.

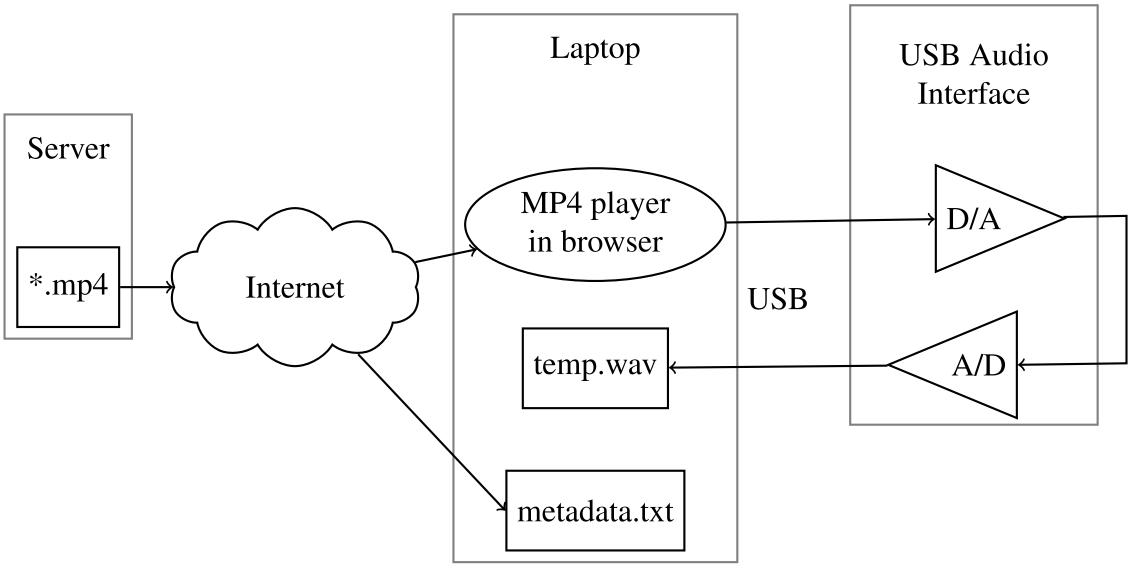

Figure 1. Audio and metadata capture path.

Different impairment sources produce different sounds. Even casual listening makes it very clear that the impairments identified in the analyses which follow are due to the original recording process and *cannot* be attributed to encoding, the network transmission, or the play and re-record processes.

Even so, we did take some steps to formalize this observation. We tested the transparency of the play-record process shown in Figure 1 by playing some locally generated uncompressed audio files and then comparing each original (played) file with the new (recorded) file. We found results consistent with the specifications of the high quality USB audio interface (Focusrite Scarlett 2i2) that we used; differences between the two files were very small compared to the impairments discussed in this paper.

We also inspected the metadata in the multimedia (.mp4) containers and found that the final audio encoder (i.e., the presentation encoder) was some form of AAC in 99% of the presentations. The distribution of presentation audio encoding bit rates is shown in Figure 2. This figure indicates that the most common rates are those between 64 and 128 kbps, followed by those between 128 and 256 kbps. The distribution of sample rates is also shown in Figure 2. Sample rates of 48 or 44.1 kHz (supporting fullband audio) were used in 31.6 or 32.1% of the cases, respectively. The rate 32 kHz (supporting super-wideband audio) was used in 34.8% of the cases. The rates 24, 22.05, or 16 kHz (supporting wideband audio) were used in 0.6, 0.2, and 0.8% of the cases respectively. These encoding conditions can provide very good quality speech. Impairments discussed in this paper are not due to the presentation encoding.

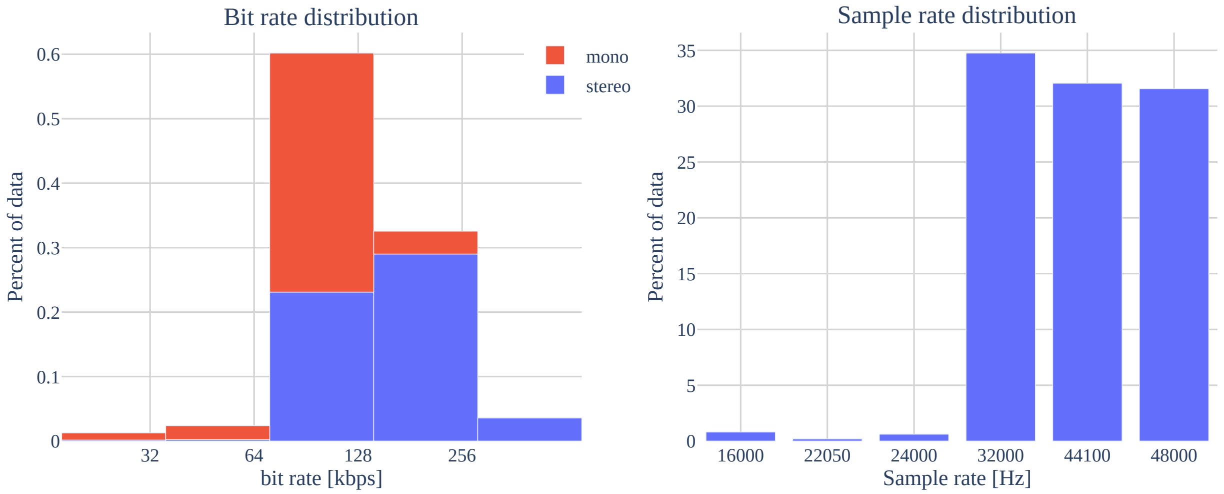

Figure 2. Bit rate distribution (left) and sample rate distribution (right).

We also considered the unlikely possibility that packet loss and jitter between the server and the laptop might impair the audio playout. This is unlikely because the rate required to stream the .mp4 files (most often 0.5 to 1.5 Mbps) is far below the measured connection bandwidth (typically 800 Mbps). In addition, the playout is non-real time and this allows for generous buffering that can mitigate the effects of any packet loss or jitter.

As an additional precaution we collected each excerpt two times. We applied seven different NR tools (see Sec. 4) to each pair of excerpts and compared the results. If any of the seven differences exceeded more than 1% of the full range of that NR tool, we excluded that excerpt from further analysis.

Between these tests and our listening experiences, we are confident that impairments identified in the following can be attributed to the original recording process — they are not associated with presentation encoding, network transmission, or play and re-record processes.

While presentation encodings are not creating significant impairments, other encodings are sometimes clearly audible in the excerpts. One explanation for these encoding artifacts is as follows. The conference organizers suggested that creating and recording an online meeting would be one pathway to a presentation recording. An online meeting can easily introduce lower rate and lower quality encodings, and, depending on the exact configuration, it can also introduce packet losses that cannot be fully concealed due to the near real-time goal of online meetings.

Our analyses cover one excerpt from each of 2592 different presentations. Random sampling and listening to the excerpts suggests that both very high quality and badly impaired presentations are common. In some cases these impairments reduce speech intelligibility and thus defeat the tutorial nature of a presentation, sometimes they increase the listening effort required, and they often make listening an unpleasant experience. Identifying impaired recordings and asking producers to submit improved versions would benefit everyone involved in a conference like this, and the associated costs would be negligible.

3 SIGNAL ANALYSES

We can analyze the signals in the temporary audio files and use generalized prior knowledge to detect impaired signals without access to any clean versions of the signals. In this section we summarize intuitive, simple signal processing measurements that identify issues with volume, dynamic range, spectral balance, clipping, and stereo images.

3.1 Channel Power

We measured the relative power of the N samples y_i in each channel of each temporary audio file as $10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2\right)$. The distribution of powers (per-channel) is shown on the left in Figure 3. A listener can typically adjust a volume control to bring presentations that are significantly above or below the norm to the desired listening level. But 45 dB of gain is required to bring the lowest signal (-67.5 dB) up to the most common level (-22.5 dB); this gain boosts the noise floor significantly, and can be dangerous when moving on to a different presentation. For 96% of the recordings the measured power level is within 15 dB of the most common value (-37.5 to -7.5 dB). The remaining 4% have lower power, which places them at some disadvantage.

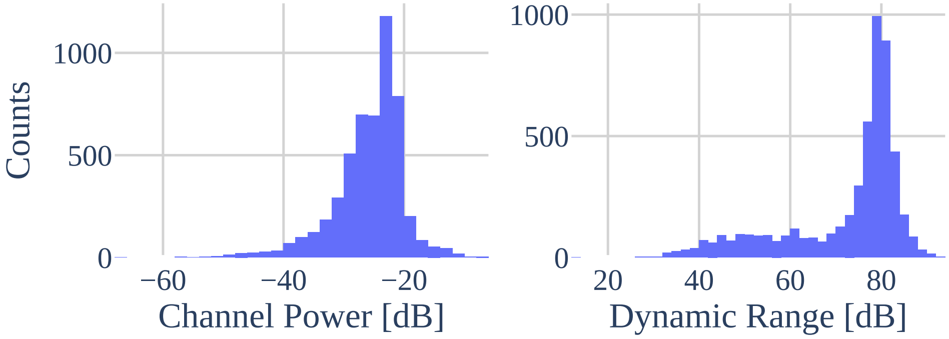

Figure 3. Distributions of relative power and dynamic range.

Eight of the original presentations had no signal in one channel or the other. For consistency, we exclude these from further analysis. Among the remaining presentations, the channel power difference never exceeds 12 dB, and it exceeds 3 dB in only 12 cases, placing the presenter comfortably within a typical stereo image.

3.2 Dynamic Range

We similarly calculated the power for each 15 ms frame in every temporary audio file. These 15 ms frames were spaced with a frame-stride of 7.5 ms and these values are consistent with the range of values often used for frame-based analysis of speech signals [27]. The dynamic range (DR) of the temporary audio file is the difference between the maximum and minimum frame power, and is shown on the right in Figure 3. DR tells how far the speech signal may be above the noise, reverberation, or other artifacts that are present between speech segments, so presentation excerpts

Table 1. Spectral balance bands and reference values.

Band	Band Limits (Hz)	Relative Power (dB)
1	[20, 50)	-20.9
2	[50, 200)	-4.9
3	[200, 350)	-3.9
4	[350, 750)	0.0
5	[750, 2000)	-6.3
6	[2000, 4000)	-12.1
7	[4000, 8000)	-12.5
8	[8000, 12,000)	-15.4
9	[12,000, 16,000)	-27.4
10	[16,000, 20,000]	-38.4

with larger DR sound better. About 75% of the excerpts have very good DR values that are in the mode centered on 80 dB. Excerpts with DRs between 40 and 65 dB could sound impaired, depending on the exact cause of the DR reduction. About 2.5% of the excerpts have DR below 40 dB (the worst DR is 12.6 dB) and these contain clearly audible and objectionable impairments.

3.3 Spectral Balance

For each excerpt we calculated the relative power in the 10 frequency bands shown in Table 1. These bands were selected to efficiently capture the spectral signature of speech and the different spectral regions that drive human perception of speech, as well as the spectral regions associated with different speech coding bandwidths. We compared the values for the excerpts with reference results that we calculated identically using studio-recorded speech and averaged over 13 minutes from 13 talkers, sampled at 48 kHz. These comparisons allowed us to find excerpts with severe spectral imbalance or coloration. For example, the reference has peak power in the band that covers 350 to 750 Hz. When we examine excerpts that are at least 10 dB below the reference in this band, we find 10 very muffled, boomy sounding excerpts (with peak power between 50 and 350 Hz) and 11 extremely tinny or thin sounding excerpts (with peak power between 750 Hz and 4 kHz). So this simple example with very generous thresholds identifies 21 of the 2592 excerpts (nearly 1%) as ones that could be dramatically improved with respect to spectral balance. Other bands and thresholds are also useful for detecting objectionable levels of additional specific spectral imbalances. Figure 4 allows one to compare relative powers for three example excerpts with the reference values of relative power. Each of these excerpts has very poor spectral balance and is extremely unpleasant to listen to.

3.4 Clipping

When speech signal levels are outside the limits that hardware or software accepts, clipping or hard limiting and associated artifacts may result. These can reduce speech quality dramatically and can be easily detected by considering the extreme values of the time-domain samples. We observed that in clean speech, around 0.1% of the samples have a value that is beyond 90% of the

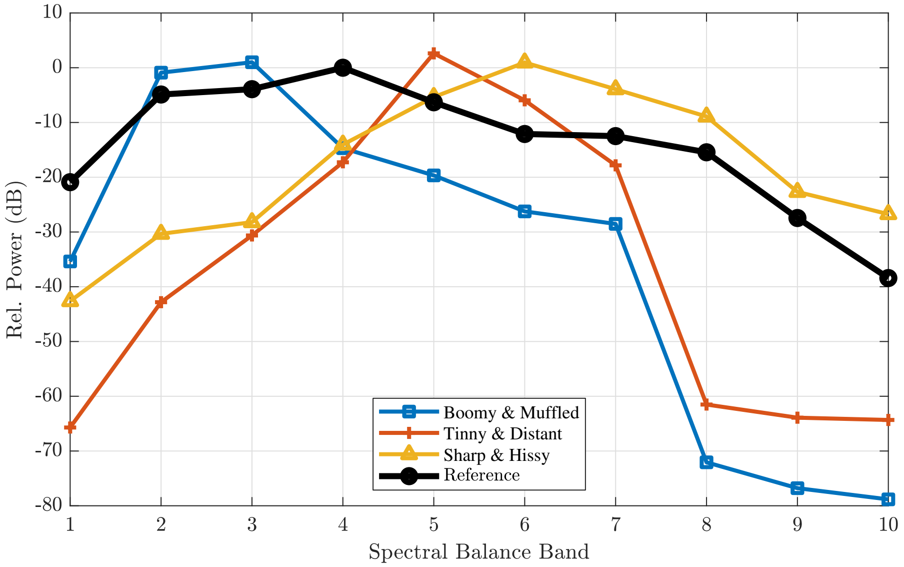

Figure 4. Three examples of excerpts with very poor spectral balance (along with reference).

full amplitude range, as determined by peak signal excursions. For example, if the nominal peak signal excursion range is ±1.0 then only 0.1% of the samples are outside the range ±0.9. If more samples are outside of this range, that is an indication that some hard limiting may be occurring, and when many samples are outside this range, clipping is likely.

We calculated the percentage of samples exceeding 90% of "full scale" in all excerpts. We compared with a threshold set at 1.0% which is 10 times the observed rate of excursion in clean speech $(10 \times 0.1\%)$. Even with this very forgiving threshold, we found nine excerpts with hard limiting and associated annoying distortions.

3.5 Channel Correlation

We also calculated the correlation between the left and right channel waveforms. For 99% of the excerpts the correlation is greater than 0.8, consistent with natural-sounding stereo or mono content. We listened to the remaining 1% and found unpleasant and highly unstable sound fields, with the source (or some spectral components of the source) often dancing between the channels. The impairment appears to be software produced, and it most commonly appears when the narration is provided by a text-to-speech (TTS) system. We also heard intermittent microphone connections and the case where the narrator appears mostly in one channel and noise dominates the other channel. And in one case a phase inversion in the recording path produced an exact -0.98 correlation and the accompanying soundfield with a very dramatic "hole-in-the-middle."

4 SPEECH QUALITY ESTIMATION

4.1 No-Reference Tools

Recently many NR speech quality tools have been developed that demonstrate high correlations with training targets. Many of these employ multi-task learning to hit multiple targets at once, which tends to improve the overall performance of all estimates. These targets can either be subjective scores from listening experiments, or objective scores from FR quality estimators. While these tools are very effective on the data they have been trained on, we can now investigate how they perform on real-world audio where no reference signal and no truth data are present.

We identified three sets of tools to use in our analyses, all of which offer estimations of overall quality via mean opinion score (MOS). We ran two flavors of NISQA: v2.0 of the model with five prediction outputs (MOS, noisiness (NOI), coloration (COL), discontinuity (DIS), and loudness (LOUD)) **mittag2021nisqa** and Baseline 2 from the ConferencingSpeech 2022 challenge [25], which outputs only MOS. We also used WAWEnets [18] in a mode that outputs four predictions of subjective scores (MOS, NOI, COL, and DIS) and seven predictions of objective scores (WB-PESQ [28], POLQA [29], STOI [30], PEMO [31], ViSQOL3 [32], ESTOI [33], and SIIB-Gauss [34]). And we used TorchAudio-Squim [24], which currently provides a tool for subjective MOS estimations using a non-matching reference and a tool for estimates of three objective targets (STOI, WB-PESQ, SI-SDR [35]). The TorchAudio-Squim subjective tool requires a non-matching reference, but specifies that any clean speech may be used as a non-matching reference. We opted to use four different non-matching references in order to characterize sensitivity to the choice of reference and ideally obtain a better MOS estimate. The non-matching references came from reference audio from a NISQA dataset **mittag2021nisqa** and an internal dataset. We selected one male and one female talker from each.

4.2 Selecting Metrics

We needed to select a smaller subset of metrics for cross-tool analysis. We opted to focus on speech quality estimates, as they are common to all three tools and should respond to all the impairments that may be present. One could make the case that intelligibility would be a better indicator of successful communication. But in this case attendees have a vast number of presentations to sift through so high listening effort or an unpleasant listening experience could easily remove a recording from consideration. Quality captures both of those elements and we selected two quality estimates from each tool.

For NISQA we choose the MOS estimates from v2.0 as well as the MOS estimate from Baseline 2 of the ConferencingSpeech 2022 challenge. We label these as NISQA MOS and NISQA22 MOS respectively. For WAWEnets we chose its MOS and PESQ estimates and label them as WAWEnets MOS and WAWEnets PESQ.

TorchAudio-Squim produced an estimate of MOS for each non-matching reference. For three references those values were very similar with correlations above 0.9. The Source 1 Female results had 0.89 correlation with the Source 1 Male results, but only 0.68 and 0.81 correlation with results

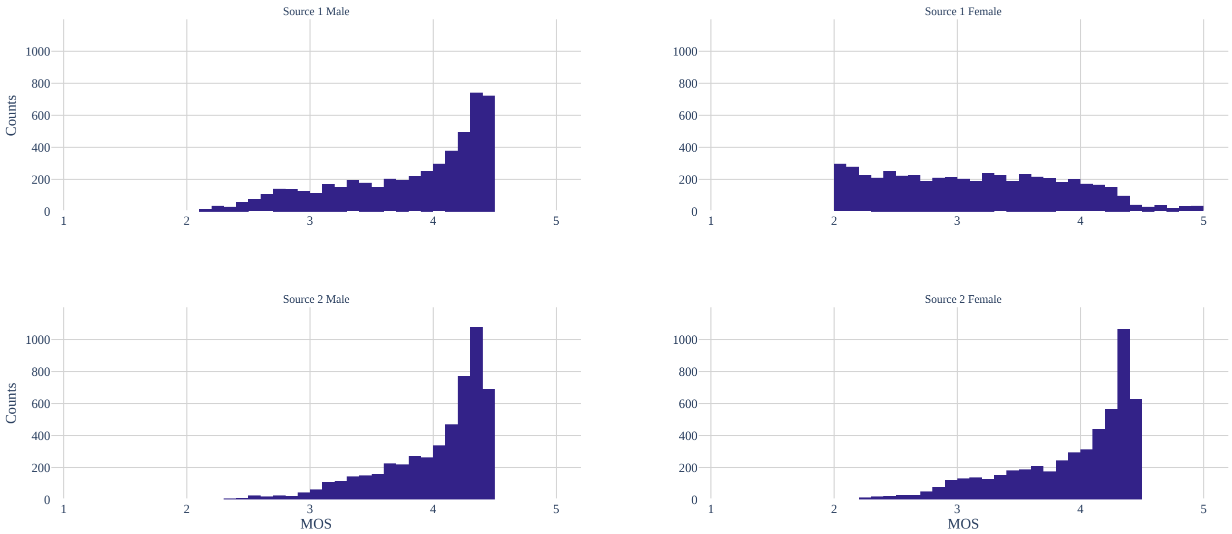

Figure 5. Per-channel distributions of Squim MOS estimations of conference speech quality with different non-matching references.

Table 2. Correlations between TorchAudio-Squim estimates across different non-matching references.

File	Source 1 Male	Source 1 Female	Source 2 Male	Source 2 Female
Source 1 Male	1.00	0.89	0.93	0.99
Source 1 Female	0.89	1.00	0.68	0.81
Source 2 Male	0.93	0.68	1.00	0.98
Source 2 Female	0.99	0.81	0.98	1.00

from Source 2, as seen in Table 2. The distributions of scores for each reference can be seen in Figure 5, and Source 1 Female again stands out in terms of distribution. To validate that all four of the non-matching references are high quality and that Source 1 Female is not unique, we evaluated the four references with all the NR MOS estimators we considered, including TorchAudio-Squim with each non-matching reference. These results are shown in Table 3, and demonstrate that all four non-matching references are high quality. In particular, Source 1 Female has the second highest average quality, so it should be a suitable non-matching reference despite its apparent disagreement with the other non-matching references.

Table 3. NR quality estimates for the selected non-matching references.

File	NISQA MOS	NISQA22 MOS	WAWEnets MOS	Source 1 Male	Source 1 Female	Source 2 Male	Source 2 Female	Average
Source 1 Male	5.01	4.80	4.47	4.45	4.46	4.30	4.38	4.55
Source 1 Female	4.87	4.68	4.45	4.46	4.47	4.36	4.42	4.53
Source 2 Male	4.29	4.35	3.78	4.39	3.59	4.46	4.44	4.19
Source 2 Female	4.33	4.47	4.04	4.41	3.71	4.44	4.45	4.26

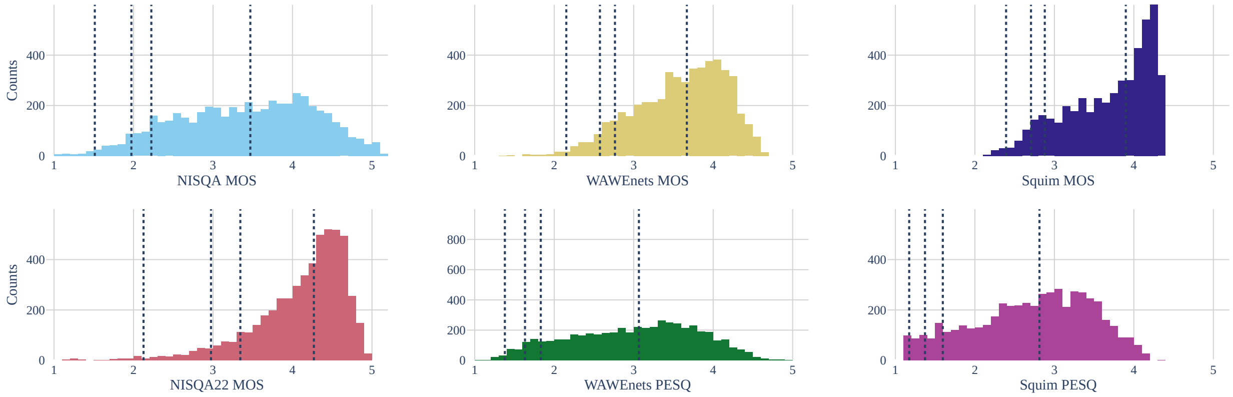

Figure 6. Per-channel distributions of conference speech quality estimates across selected metrics. Dotted lines from left to right denote the 1st, 5th, 10th, and 50th percentiles respectively.

Table 4. Correlations between NR speech quality estimates.

	NISQA MOS	NISQA22 MOS	WAWEnets MOS	WAWEnets PESQ	Squim MOS	Squim PESQ
NISQA MOS	1.00	0.73	0.67	0.65	0.50	0.60
NISQA22 MOS	0.73	1.00	0.71	0.63	0.56	0.60
WAWEnets MOS	0.67	0.71	1.00	0.94	0.61	0.79
WAWEnets PESQ	0.65	0.63	0.94	1.00	0.58	0.83
Squim MOS	0.50	0.56	0.61	0.58	1.00	0.54
Squim PESQ	0.60	0.60	0.79	0.83	0.54	1.00

To minimize the number of metrics and still treat all data fairly, we opted to average the results from all four non-matching references into a single score we label as Squim MOS. Averaging over a sample from the population of high quality non-matching references yields a more confident MOS estimate than that of a single non-matching reference. This is a natural extension to estimating audio quality by averaging scores from multiple listeners. In fact, compared with the four individual TorchAudio-Squim MOS estimates, this average MOS is equally or better correlated to the other quality metrics under consideration. We also used the PESQ estimate from the objective TorchAudio-Squim tool, and label it as Squim PESQ. Distributions of the selected metrics are shown in Figure 6, and correlations are given in Table 4.

4.3 Reconciling Different Metrics

In general, the different metrics do not seem to yield compatible results. They have a wide range of correlations and do not consistently rank the excerpts. There are no excerpts that are in the bottom 1% for all six metrics and only 7 excerpts are represented in the bottom 5% of all metrics. But here, the task is to simply identify low quality in real world recordings, which is significantly different from the MOS correlation paradigm that often drives tool development in the lab. While correlations are focused on the global ranking across all files, if the metrics are ranking files differently but in general are all successfully identifying lower-quality speech, then they are succeeding at our task. Listening to the excerpts in the low-quality tail of any metric confirms that all six reliably identify low quality speech.

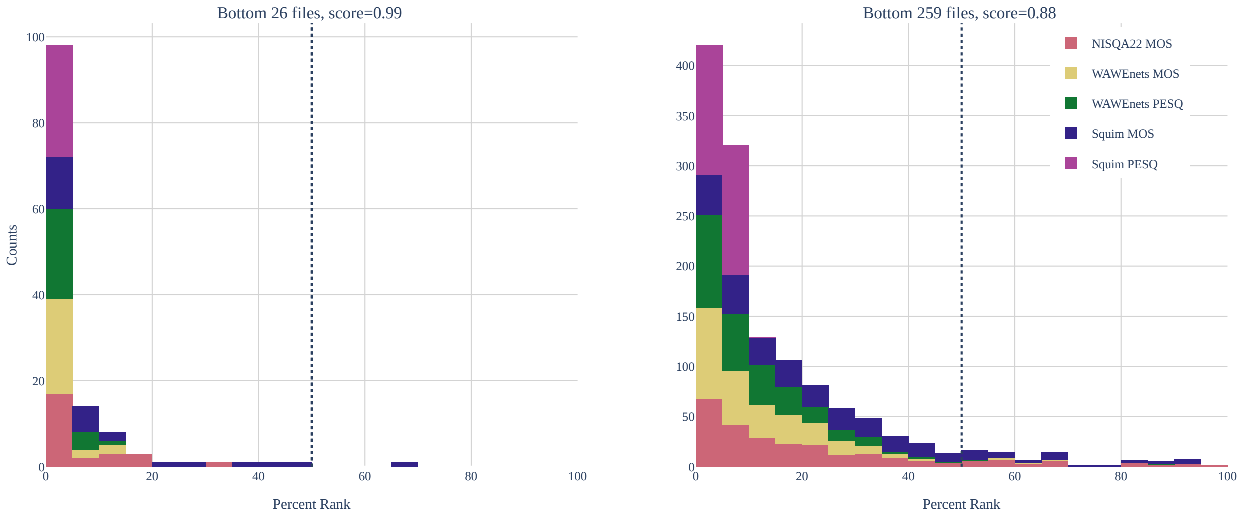

Figure 7. Stacked histogram demonstrating the consistency of NISQA MOS with every other metric. The relative ranks for each comparison metric are shown for the bottom k excerpts of the NISQA MOS distribution. An excerpt is considered consistently labelled between NISQA MOS and a comparison metric when its percent rank is below 50%.

This motivates a different measure of agreement between metrics. We focus on the low tails of a given metric's distribution and then ask every other metric a binary question: Do the excerpts in those tails have higher or lower quality than the median score in your own distribution?

To make this explicit, we define sets of scores for all 2592 excerpts, across the metrics we consider in this paper as $X_{i,j}$, with $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, such that $X_{i,j}$ represents the score from metric i on excerpt j with $m = 6$ and $n = 2592$. Note that for this analysis we average scores over both channels in each excerpt. We then use each metric individually to rank all excerpts. This allows us to identify the excerpts with the lowest k scores for metric i as $\{r_1, r_2, \ldots, r_k\}$ such that $X_{i,r_1} = X_{i,(1)}$, where $X_{i,(1)}$ is the first order statistic, and so on. Let $p_{i,50}$ denote the 50th percentile value for metric i. Then we can define a consistency measure $C(i,k)$ as the fraction of excerpts with the lowest k scores from metric i that are in the bottom half of a comparison metric's distribution averaged over all comparison metrics. In particular

$$C(i,k) = \frac{1}{m-1} \sum_{j \neq i} \frac{1}{k} \sum_{n=1}^{k} \mathbb{1}[X_{j,r_n} \leq p_{j,50}], \tag{1}$$

where $\mathbb{1}$ is the indicator function. Figure 7 shows an example of this focusing on NISQA MOS with $k = 26$ (lowest 1% of the data) and $k = 259$ (lowest 10%). Table 5 shows the consistency

Table 5. Consistency scores for all metrics across a range of selected excerpts; t is the percentile which determines the number of excerpts k.

t	NISQA MOS	NISQA22 MOS	Squim MOS	Squim PESQ	WAWEnets MOS	WAWEnets PESQ
1%	0.99	1.00	0.98	0.98	1.00	0.98
5%	0.90	0.97	0.94	0.95	0.98	0.97
10%	0.88	0.95	0.90	0.93	0.96	0.95

Table 6. Correlations between NISQA subjective estimates.

	MOS	NOI	DIS	COL	LOUD
MOS	1.00	0.64	0.84	0.88	0.78
NOI	0.64	1.00	0.41	0.46	0.56
DIS	0.84	0.41	1.00	0.80	0.53
COL	0.88	0.46	0.80	1.00	0.73
LOUD	0.78	0.56	0.53	0.73	1.00

results over a variety of values of k, presented as percentages of the total number of excerpts. From this it is clear that each metric identifies excerpts that every other metric categorizes as lower quality.

4.4 Individual Tool Analyses

Both NISQA and WAWEnets offer insight into multiple dimensions of speech quality (NOI, DIS, and COL), as well as overall quality estimations (MOS). These additional dimensions provide insight into what sort of impairments are driving the overall quality in the conference presentation excerpts. The NISQA correlations between MOS and its other components are 0.64 (NOI), 0.84 (DIS), and 0.88 (COL), as can be seen in Table 6. WAWEnets reports very similar correlations between MOS and those dimensions with 0.70 (NOI), 0.82 (DIS), and 0.92 (COL), which can be seen in Table 7. Note that the ranking among all three additional components is identical among the tools. Listening to the low tails of each of these dimensions confirmed that each dimension identified low quality speech with different types of impairments. For example, low NOI scores identified overwhelming background noise, low DIS scores were associated with noise suppression and codec effects, and low COL scores found extremely muffled or tinny audio.

Table 7. Correlations between WAWEnets subjective estimates.

	MOS	NOI	DIS	COL
MOS	1.00	0.70	0.82	0.92
NOI	0.70	1.00	0.37	0.40
DIS	0.82	0.37	1.00	0.78
COL	0.92	0.40	0.78	1.00

The most obvious potential weakness of TorchAudio-Squim is in its MOS prediction and the reliance on a non-matching reference. Figure 5 shows the distribution of values can be very dependent on the selection of a non-matching reference, although the values for the other three references were very well correlated (values of 0.98, 0.94, and 0.99). Figure 5 also shows some apparent boundaries on output values. No values less than 2 were observed and three of the non-matching references produced hard upper limits near 4.5, while the limit for the other reference was 5.0. These apparent constraints could limit TorchAudio-Squim's ability to produce reliable MOS estimates for very high and low quality speech.

5 RECOMMENDATIONS AND CONCLUSION

We have analyzed 2592 audio excerpts from real-world recorded presentations from a recent conference. Those presentations cover a vast range of recording environments and they provide a unique opportunity for the application of recent NR speech quality tools and other signal analyses. Now we present some simple considerations for improved live and recorded speech for meetings and conferences.

In Section 3 we offered some simple example signal measurements, tested them against very lenient thresholds, and easily identified a total of 116 excerpts that are badly impaired with respect to dynamic range, spectral balance, clipping, or stereo image. In Section 4 we discussed speech quality metrics that successfully identify impaired speech in the tails of their distributions, with the 1%, 5%, 10%, and 50% thresholds for the six metrics shown in Figure 6. We focus again on the six quality metrics from Section 4.2 and consider the 26 excerpts that comprise the lowest 1% for each metric. Together, the six metrics identify 88 excerpts. This is more than the 26 excerpts that would result from perfect agreement and less than the 156 (6×26) excerpts that would result from perfect disagreement.

Of these 88 impaired excerpts, 25 were also identified through signal analysis, as discussed in Section 3, and the other 63 are unique. This emphasizes that speech quality is not fully characterized by those basic signal analyses, especially when digital speech coding and noise suppression artifacts are present. Together, the two sets of tools identify 179 highly impaired excerpts, which is 7% of the total. We have listened to all of the identified excerpts and all have much room for improvement. Unsuppressed noise, high levels of noise suppression artifacts (tonal artifacts, spectral modulation of speech, gating of speech), poor spectral balance and excessive resonances (muffled, thin, distant, tinny, boomy, ringing), excessive reverberation, low-rate speech coding artifacts, hard limiting, clipping, and other impairments appear alone and in combinations.

Meeting and conference experiences could be improved if organizers would identify impaired audio and work with contributors to resolve issues. Impairments commonly heard in the dataset can often be reduced or eliminated with minimal effort and expense as follows.

- Select the quietest available recording location and avoid locations with excessive reverberation.

- Adjust hardware and software level controls to the highest level that does not cause excessive clipping or hard limiting.

- External microphones (even some inexpensive ones) can have significant advantages over built-in microphones. They can be optimally positioned for less noise and reverberation and to capture better spectral balance.

- Use the minimal amount of noise suppression required. Overly aggressive noise suppression can create prominent artifacts, cause unstable speech levels and spectral balance, and can cause gating of the speech.

- Some current TTS systems can produce high-quality natural-sounding speech but other TTS systems produce very unnatural sounding speech that is difficult to listen to. As a separate issue, when TTS is used, the TTS output should connected directly to the input of the recording system. Expecting the microphone on a recording system to properly capture the sound from the speaker of the TTS device is unreasonable and can result in high noise and extremely poor spectral balance, and this was observed in multiple presentations.
- Record locally with recording software if at all possible. Setting up a "meeting" and recording it is likely to introduce significant additional impairments including lower rate speech coding and packet loss.

Our work shows that real-world conference speech is often impaired, several available tools automatically and consistently detect impairments, and audio impairments are often easily mitigated.

REFERENCES

[1] J. Liang and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Vehicular Technology Conference*, vol. 3, Jun. 1994, pp. 1719–1723. DOI: `10.1109/VETEC.1994.345390`.

[2] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006. DOI: `10.1109/TASL.2006.883177`.

[3] D. Kim and A. Tarraf, "ANIQUE+: A new American National Standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, Spring 2007. DOI: `10.1002/bltj.20228`.

[4] T. H. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006. DOI: `10.1109/TASL.2006.883253`.

[5] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *Proc. European Signal Processing Conference*, Nov. 2016, pp. 2315–2319. DOI: `10.1109/EUSIPCO.2016.7760662`.

[6] M. Hakami and W. B. Kleijn, "Machine learning based non-intrusive quality estimation with an augmented feature set," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2017, pp. 5105–5109. DOI: `10.1109/ICASSP.2017.7953129`.

[7] J. Ooster and B. T. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, May 2019, pp. 636–640. DOI: `10.1109/ICASSP.2019.8682754`.

[8] J. F. Santos and T. H. Falk, "Towards the development of a non-intrusive objective quality measure for DNN-enhanced speech," in *Proc. Eleventh Intl. Conf. on Quality of Multimedia Experience*, 2019, pp. 1–6. DOI: `10.1109/QoMEX.2019.8743156`.

[9] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2019, pp. 85–89. DOI: `10.1109/WASPAA.2019.8937202`.

[10] G. Mittag, R. Cutler, Y. Hosseinkashi, *et al.*, "DNN no-reference PSTN speech quality prediction," in *Proc. Interspeech*, Oct. 2020, pp. 2867–2871. DOI: `10.21437/interspeech.2020-2760`.

[11] M. Liu, J. Wang, W. Yi, and F. Liu, "Neural network-based non-intrusive speech quality assessment using attention pooling function," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 20, May 2021. DOI: `10.1186/s13636-021-00209-4`.

[12] J. Serrà, J. Pons, and S. Pascual, "SESQA: Semi-supervised learning for speech quality assessment," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 381–385. DOI: `10.1109/ICASSP39728.2021.9414052`.

[13] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. 2021 IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 391–395. DOI: 10.1109/ICASSP39728.2021.9413877.

[14] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2022, pp. 886–890. DOI: 10.1109/ICASSP43922.2022.9746108.

[15] A. Ragano, E. Benetos, and A. Hines, "More for less: Non-intrusive speech quality assessment with limited annotations," in *Proc. Thirteenth Intl. Conf. on Quality of Multimedia Experience*, 2021, pp. 103–108. DOI: 10.1109/QoMEX51781.2021.9465410.

[16] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131. DOI: 10.21437/Interspeech.2021-299.

[17] N. Nessler, M. Cernak, P. Prandoni, and P. Mainar, "Non-intrusive speech quality assessment with transfer learning and subject-specific scaling," in *Proc. Interspeech*, 2021, pp. 2406–2410. DOI: 10.21437/Interspeech.2021-1685.

[18] A. A. Catellier and S. D. Voran, "WAWEnets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 331–335. DOI: 10.1109/ICASSP40776.2020.9054204.

[19] X. Jia and D. Li, "A deep learning-based time-domain approach for non-intrusive speech quality assessment," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 477–481.

[20] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 316–320. DOI: 10.1109/ICASSP39728.2021.9414182.

[21] M. Yu, C. Zhang, Y. Xu, S.-X. Zhang, and D. Yu, "MetricNet: Towards Improved Modeling For Non-Intrusive Speech Quality Assessment," in *Proc. Interspeech*, 2021, pp. 2142–2146. DOI: 10.21437/Interspeech.2021-659.

[22] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 911–915. DOI: 10.1109/ICASSP40776.2020.9053366.

[23] R. . S.-W. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023. DOI: 10.1109/TASLP.2022.3205757.

[24] A. Kumar, K. Tan, Z. Ni, *et al.*, “TorchAudio-Squim: Reference-less speech quality and intelligibility measures in TorchAudio,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096680.

[25] G. Yi, W. Xiao, Y. Xiao, *et al.*, “ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications,” in *Proc. Interspeech 2022*, 2022, pp. 3308–3312. DOI: 10.21437/Interspeech.2022-10597.

[26] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, “The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments,” *Computer Speech & Language*, vol. 26, no. 1, pp. 52–66, 2012. DOI: https://doi.org/10.1016/j.csl.2010.12.003.

[27] L. R. R. Schafer, *Digital Processing of Speech Signals*. Upper Saddle River, NJ: Prentice Hall, 1978.

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

[29] *ITU-T Recommendation P.863, perceptual objective listening quality prediction*, Geneva, 2018.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, ISSN: 1558-7916. DOI: 10.1109/TASL.2011.2114881.

[31] R. Huber and B. Kollmeier, “PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006. DOI: 10.1109/TASL.2006.883259.

[32] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “ViSQOL v3: An open source production ready objective speech and audio metric,” in *Proc. Twelfth International Conference on Quality of Multimedia Experience*, 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123150.

[33] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. DOI: 10.1109/TASLP.2016.2585878.

[34] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility metric based on information theory,” *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018. DOI: 10.1109/LSP.2017.2774250.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630. DOI: 10.1109/ICASSP.2019.8683855.

Form NTIA-29
(6/04)

U.S. Department of Commerce
National Telecommunications and Information Administration

BIBILIOGRAPHIC DATA SHEET

1. Publication Number
TM-24-571

2. Government Accession Number

3. Recipient's Accession Number

4. Title and Subtitle
Joint Analyses of No-Reference Speech Quality Estimation Tools and Conference Speech Recorded in Diverse Real-World Conditions

5. Publication Date
July 8, 2024

6. Performing Organization Code
NTIA/ITS.P

7. Author(s)
Jaden Pieper and Stephen D. Voran

9. Project/Task/Work Unit No.
3142012-300

8. Performing Organization Name and Address
National Telecommunications and Information Administration
Institute for Telecommunication Sciences, 325 Broadway, Boulder, CO 80305

10. Contract/Grant Number

11. Sponsoring Organization Name and Address
National Telecommunications and Information Administration
Herbert C. Hoover Bldg. 14th & Constitution Ave., NW,
Washington, DC 20230

12. Type of Report and Period Covered

14. Supplementary Notes

15. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

Recently, prerecorded audio and video presentations, as well as virtual meetings, have become a common component of professional life, due to health and environmental considerations. This places new responsibility on participants to generate audio that is of sufficiently high quality to effectively communicate. This memorandum provides analyses of real world audio from a virtual component of a 2023 conference which encompasses a wide range of recording environments and conditions. We use both signal analyses and novel machine learning based no-reference speech quality estimators and we evaluate their performance relative to each other. We utilized NISQA, WAWEnets, and TorchAudio-Squim, and found that while their scores show only modest agreement, we can use each to successfully identify low-quality speech. Finally we offer remediation steps for speech conferencing to avoid many of the impairments observed in this work.

16. Key Words *(Alphabetical order, separated by semicolons)*

conference speech, intelligibility, no-reference speech quality assessment, speech impairments, speech quality

17. Availability Statement

☑ Unlimited

□ For Official Distribution

18. Security Class. *(This report)*
Unclassified

20. Number of Pages
26

19. Security Class. *(This page)*
Unclassified

21. Price
N/A

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.gov.