

Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech

Erica Cooper, Junichi Yamagishi

National Institute of Informatics, Japan

ecooper@nii.ac.jp, jyamagis@nii.ac.jp

Abstract

Mean Opinion Score (MOS) is a popular measure for evaluating synthesized speech. However, the scores obtained in MOS tests are heavily dependent upon many contextual factors. One such factor is the overall range of quality of the samples presented in the test – listeners tend to try to use the entire range of scoring options available to them regardless of this, a phenomenon which is known as *range-equalizing bias*. In this paper, we systematically investigate the effects of range-equalizing bias on MOS tests for synthesized speech by conducting a series of listening tests in which we progressively “zoom in” on a smaller number of systems in the higher-quality range. This allows us to better understand and quantify the effects of range-equalizing bias in MOS tests.

Index Terms: mean opinion score, subjective evaluation, listening tests, range-equalizing bias

1. Introduction

Listening tests where human listeners rate the quality of generated samples are the gold standard for evaluating models of synthesized speech such as text-to-speech or voice conversion, and the Mean Opinion Score (MOS) test [1], where listeners rate audio samples one by one on an ordinal scale, is a popular testing paradigm for rating and comparing different approaches. However, MOS tests are known to be very context-dependent, where the context includes the demographics of the listeners that were recruited to participate in the test, their personal opinions and individual differences, the environment in which they are listening to samples, the types of synthesis systems included in the test, the sentences chosen to synthesize test material, and even the appearance of the testing interface. In this paper we investigate the effect of *range equalizing bias*, which is the tendency of listeners to use the entire range of choices on the rating scale, regardless of the absolute quality of the samples present in the test. To do this, we start with an existing large-scale database of audio samples and their MOS ratings, and conduct new listening tests where we progressively “zoom in” on the high-quality region of synthesis systems. By quantifying the ways in which MOS ratings change as listening test context is changed systematically, we can better understand the effects of context on listening test results, which can help researchers to design their listening tests in more principled ways, and which can also help to improve automatic MOS prediction.

The BVCC dataset [2] is a large-scale dataset of MOS ratings for samples from 187 different text-to-speech and voice conversion systems covering over a decade of progress in these fields. These samples were evaluated all together in one large-scale MOS test. This dataset provided for the first time a huge variety of synthesized samples all evaluated together in

the same context, and has since been used as training material for automatic MOS predictors [3, 4]. However, this dataset has several shortcomings. The wide variety of synthesis methods present in the dataset provide good coverage and generalizability [3], but this also means that there are fewer fine-grained and significant differences to be found between synthesizers of similar quality, and that the ability of MOS predictors trained on this data to correctly rank synthesis systems that are close in quality will be impaired. The number of systems is also much larger than a typical or representative listening test. Furthermore, present-day researchers are less likely to want to compare their proposed synthesis methods with ones from many years ago. Present-day, high-quality synthesis systems are the more relevant comparisons, and there is especially room for improvement in this regard as shown from the results of the VoiceMOS Challenge – while low-quality samples were consistently easy to predict for all teams, middle-range and high-quality systems remained more difficult to predict accurately. These points motivated us to revisit the BVCC dataset and conduct new listening tests on progressively smaller subsets of the synthesis systems, focusing in towards the top of the range of quality. Data from these new listening tests allow us to observe and quantify the effects of range-equalizing bias on MOS of synthesized speech. We expect to see scores for some systems to decrease as listeners attempt to use the entire range of score choices, and we also expect to be able to find more significant differences between systems.

The distributions of scores reveal the presence of range-equalizing bias in MOS listening tests for synthesized speech, while the skew of the distributions demonstrate that listeners nevertheless do maintain some sense of absolute quality while completing the test. We can see more significant differences emerge as the range of synthesis systems becomes smaller, and we can also observe a substantial decrease of about one MOS point in scores for some systems. In Section 2 we discuss some other studies which address context-dependency and bias in perceptual tests. In Section 3 we describe our data and listening test design. Section 4 describes the results of our listening tests and analysis of what happens when a smaller range of synthesis systems is evaluated. We conclude in Section 5.

2. Related work

A comprehensive review of biases in listening tests [5] documents the various factors that can affect listener ratings. These include recency, listener expectations, stimulus frequency, perceptually non-equal spacing between the words chosen for the rating categories that may vary by language, and even factors such as the appearance of the testing equipment or user interface. The range-equalizing bias effect is described as a “rub-

ber ruler,” meaning that listeners tend to use the entire available range of choices, regardless of “absolute” quality. The authors recommend using anchoring techniques to counteract this.

However, even listening test paradigms that use anchoring are not completely immune to this type of bias. The same authors [6] systematically varied the samples in a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) test [7], in which listeners are presented with samples from several systems including a reference sample and one or more anchor samples, finding that even when the anchors are held constant, the scores can change depending on whether there are more higher-quality or lower-quality samples present in the test.

The presentation and context of the audio sample being evaluated can affect the scores, even when the sample itself stays the same – scores can differ whether a sample under consideration is presented in isolation, or in a paragraph or dialogue context, and these scores do not even necessarily correlate with each other [8]. Scores can also be affected by the wording of the question or instructions being given to the listeners [9, 10].

Listeners may also differ in their personal preferences, and it is therefore important to collect enough different opinions. In a revisiting of the 2013 Blizzard Challenge [11] listening test data, the rank order and number of significant differences were found to change depending on sentence coverage and number of listeners, and statistical significance only stabilized at 30 listeners [12]. Another revisiting of Blizzard 2013 re-evaluated three representative systems from that challenge alongside four modern neural ones, finding that MOS ratings for the three older systems dropped a full point when evaluated together with the modern systems, even though the rankings were preserved [13]. Furthermore, a significant difference between two of the older systems was lost in the new test, indicating that their differences became “compressed” by the presence of the better systems.

These prior works emphasize that MOS is a relative rather than absolute measure, and that ratings heavily depend on many factors including the overall range of quality of the samples presented in the test. In this work, we aim to more systematically investigate and quantify how MOS changes as the overall quality of samples in the test changes, and we also aim to better understand and characterize the differences between listening tests that have a large vs. small range of qualities represented.

3. Data and listening test design

3.1. The BVCC dataset

We used audio samples from the BVCC dataset [2], a large-scale set of MOS ratings for samples from 187 different systems (including reference natural speech) from past Blizzard Challenges [14], Voice Conversion Challenges [15], and ESPnet-TTS [16]. These MOS ratings have been made public by the organizers of the VoiceMOS Challenge [4]. The corresponding audio samples are all sampled at 16 kHz, and have been amplitude-normalized using sv56 [17]. 38 samples per synthesis system were chosen for BVCC, balancing for genre (e.g., news, audiobook, conversational) and source/target speaker (for voice conversion systems) where relevant.

3.2. New listening tests

We conducted new listening tests to re-evaluate these synthesis systems in different contexts, “zooming in” on subsequently smaller, higher-quality regions of the original BVCC listening test data. We sorted the 187 systems by their MOS ratings in the original BVCC data, and created four subsets by approxi-

mately halving the number of systems each time, keeping the highest-rated half according to their original ratings. 32 samples per system were selected for all tests out of the original 38 samples, maintaining genre/speaker balance per challenge. The listening test sets (lists of samples to be rated by one listener) were designed to contain about the same number of samples regardless of the zoom level, which we accomplished by doubling the number of samples from a given system that appear in one set every time the number of systems is halved. The entire test was designed to be taken by 200 unique listeners who are native speakers of English, defined as having spoken English as their first language since birth. Our test design has fewer ratings per sample for the more zoomed-out tests and smaller numbers of listeners for the more zoomed-in tests as a necessary tradeoff due to budgetary constraints. In a pre-test questionnaire, we asked listeners for their age range, gender, whether or not they have any hearing impairment, and their dialect of English. The question asked to listeners for each sample was, “*The voice you hear is a computer-generated voice (synthesized voice). While some synthetic voices sound high quality and natural, some of them may be slightly degraded by computer processing and sound artificial, or their voices and inflections may sound discontinuous or unnatural. Please rate the overall quality of the synthesized voice subjectively on a 5-point scale from ‘very good’ to ‘very bad’.*” The rating options were *Very Bad, Bad, Fair, Good, and Very Good*, corresponding to numerical MOS ratings from 1-5. Ethics approval was obtained from our institution. Listeners were not permitted to complete more than one set from any of the zoom levels, to obtain ratings from a larger variety of listeners and to avoid cross-contamination of the different listening test contexts of each zoom level. Details of the four zoomed-in test parts, as well as the original BVCC listening test for comparison, are in Table 1. The 6% zoom test contains the top 11 systems from the original BVCC, and we note that five of these are in fact natural speech.

Table 1: *Details of listening test design. Information for the original BVCC listening test from [2, 3] is included as the 100% zoom level for comparison.*

Zoom level	Number of systems	Samples per system in a set	Number of listeners	Ratings per sample	Ratings per system
100%	187	1	304	8	304
50%	93	2	64	4	128
25%	46	4	64	8	256
12%	23	8	48	12	384
6%	11	16	24	12	384

4. Results

We can observe the change in the distributions of the ratings in the original test compared to the new tests, the emergence of more statistically-significant differences, and an overall trend of scores decreasing as we zoom in to the higher-quality region.

4.1. Distributions of scores of different listening tests

Figure 1 shows the distributions of the listener ratings for each zoomed-in new listening test (bottom row), and the ratings for the same subset of systems in the original BVCC dataset (top row) where all listeners heard samples from all 187 systems.

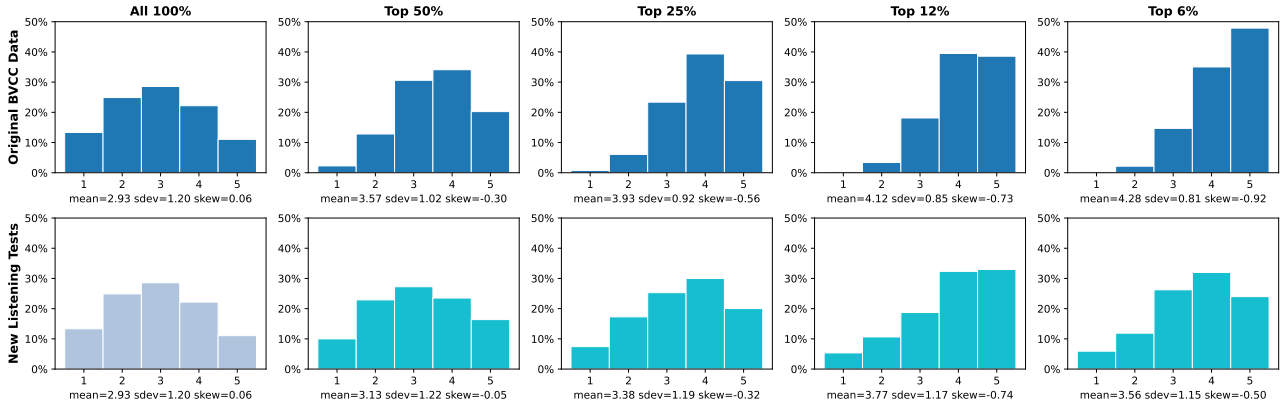


Figure 1: Distributions of listener scores for subsets of the original BVCC dataset (top), and new zoomed-in listening tests (bottom). The 100% zoom test was not repeated in our new listening tests, but we copy it in the bottom row for easier visual comparison. Bar heights show the percent of ratings with the given label.

Statistics for mean, standard deviation, and skew for each of these distributions are also shown. We can observe the presence of range-equalizing bias at the higher zoom levels – while no samples from the top 11 systems (6% zoom) received ratings of 1 in the original BVCC listening test, we nevertheless observe many “1” ratings in the new test where listeners *only* heard samples from the top 6% of systems and thus gave the worst ones a rating of 1. However, we can also observe that listeners nevertheless maintain some sense of absolute quality at the same time that they adjust their ratings to fill the whole scale – this is demonstrated by the strong skew in the distributions of the new listening tests as we zoom in towards a smaller number of higher-quality samples. Despite the fact that these listeners have not heard the samples from the lower quality systems, they still select higher scores more frequently.

4.2. Emergence of significant differences

As we zoom in towards the top of the range of quality, we expect more significant differences between systems to emerge. Table 2 shows the comparison of the number of statistically-significant differences in the relevant subsets of the original BVCC data, along with the numbers for each new test. Statistical significances were computed using the Mann-Whitney U test following [18] with Bonferroni correction for multiple comparisons. We observe that as the range of qualities of systems under evaluation becomes narrower, the number of statistical differences tends to increase as expected. This indicates that some differences between systems are being “compressed” by including a wider range of systems in the listening test. We can also observe that the number of statistically-significant pairs of systems in the top 6% of systems is 25 in both the 12% listening test and the 6% test, indicating that all of the perceivable differences may have been found – this could be further corroborated by doing a pairwise comparison between the top 6% of systems, as pairwise comparisons can be seen as an extreme case of zooming in. We leave this for future work.

4.3. Correlations

There were only two statistically-significant differences among the top 11 systems in the original BVCC listening test data, which raises the possibility that the rankings of systems may change in each test. We can measure this by looking at correla-

Table 2: Number of statistically-significant pairs of systems that were found in subsets of the original BVCC data and new listening tests. The rows are listening tests, and the columns say which top percent subset out of the 187 systems are being considered. The largest number of significant differences for a given percent subset are indicated in **bold**.

	100%	50%	25%	12%	6%
BVCC	13068	2400	354	61	2
50% test	-	1823	381	75	11
25% test	-	-	562	117	17
12% test	-	-	-	152	25
6% test	-	-	-	-	25

tions, including rank-based correlation measures. We focus on just the top 11 systems which were present in all listening tests for conciseness, and measure correlations of their scores in each new listening test with their scores in the original BVCC listening test. We report the following system-level and utterance-level correlation coefficients in Table 3: the linear correlation coefficient (LCC), which is basic correlation; Spearman’s rank correlation coefficient (SRCC), which considers how much the rankings stayed the same; and Kendall’s Tau (KTAU), another ranking-based correlation measure that is more robust to errors.

Table 3: Correlations of scores of top 11 systems in each new listening test compared to original BVCC data.

	LCC (sys)	LCC (utt)	SRCC (sys)	SRCC (utt)	KTAU (sys)	KTAU (utt)
50% test	.85	.53	.83	.52	.66	.38
25% test	.71	.38	.68	.36	.48	.26
12% test	.67	.31	.63	.29	.41	.20
6% test	.42	.20	.31	.16	.24	.11

Correlations decrease substantially as we zoom in. The two pairs of systems with significant differences in the original BVCC data do retain their order as well as their significant differences, and the systems that switch ranks are ones that did not have significant differences in the original test.

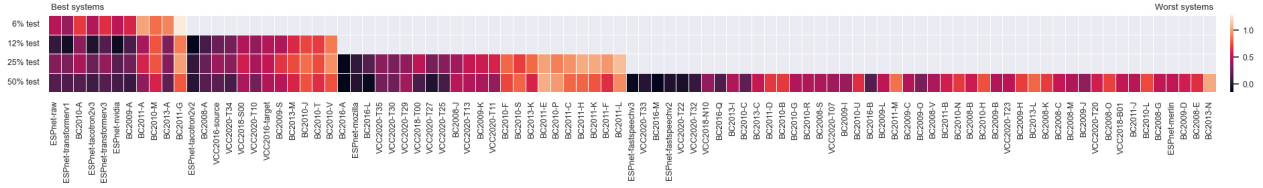


Figure 2: Heatmap of each system’s original score in the BVCC listening test, minus its score in the new tests.

4.4. Decrease in MOS ratings for some systems

In a listening test that evaluates a small number of high-quality systems, it is expected that the worst system will have a low score, even if it was rated highly in a test with a wider range of systems – that is to say, we expect some systems’ scores to decrease as we zoom in. We examine this in Figure 2, which is a heatmap showing the drop in scores of each system in each new test compared to the original BVCC data. Darker colors indicate that the score did not change much from the original listening test, and lighter colors indicate a larger score drop. Gray boxes indicate that a system was not present in a given test. Synthesis systems are sorted first according to their rank order in the 6% zoom test, then remaining systems in the 12% test are sorted according to their order in that test, and so on. It can be seen that systems that ranked low in their respective tests had larger score drops. The largest drop was of 1.28 MOS points for system BC2011-G, which was the worst-ranked system in the 6% zoom-in test. Several systems had somewhat higher scores in some of the new tests compared to BVCC; these are typically the higher-quality systems such as ESPnet. The system that had the largest increase in score (0.15 MOS points) was BC2016-A, the professional British English audiobook speech used as training data in the Blizzard Challenge 2016.

In Table 4, we report the largest drop in score for each new listening test compared to the original score in BVCC, along with the name of the synthesis system with this score drop.

Table 4: Synthesis systems with the largest drops in MOS in each new listening test compared to the original BVCC.

Test	Largest score drop	System ID
50% test	1.05	BC2011-E
25% test	1.11	BC2011-L
12% test	0.89	BC2011-G
6% test	1.28	BC2011-G

Linearly fitting score drop vs. zoom level reveals a slope close to zero, indicating that we may generally expect a system’s MOS to possibly drop by up to around an entire point if we re-evaluate it in a finer-grained context, regardless of the zoom level. Interestingly, Blizzard Challenge 2011 system G had the largest drop in MOS in both the 6% test and the 12% test, becoming the worst-ranked synthesis system in both tests, indicating that perhaps this system’s ranking in the original BVCC test results may have been unreliable.

5. Conclusions

The design of our set of listening tests that progressively “zoomed in” on the high-quality region of a large range of syn-

thesis systems shows a clear presence of range-equalizing bias. The worst system in the smallest group of top systems dropped by a full 1.28 MOS points, emphasizing the strong contextual effects of MOS-based listening tests. We can advise that if researchers want to identify more significant differences between systems that are close together in quality, then it is better to keep the overall range of quality of systems in the entire listening test fairly small. However, it is also important to consider that this may substantially lower the ratings for the worst systems in the group, even if they are overall of relatively high quality.

One limitation of this study is that when we chose the subsets of the top systems, our choices were based on the original ranking of all 187 systems in BVCC, which we found in Section 4.3 may not represent the “true” ranking of the systems. Many of the differences between systems of similar quality were not significant, and ranking order changed upon closer inspection.

We would like to conclude by observing one interesting phenomenon – the naturalness of Tacotron 2 [19] appears to have gotten *worse* over the years! The original paper reported a MOS of 4.53 in 2018 with “quality close to that of natural human speech,” whereas a study from 2020 [16] reported a MOS of 4.20, and more recent results from 2022 [20] reported a MOS for of 3.01. Of course, it’s not at all valid to make direct comparisons between MOS tests with completely different listeners, test sentences, training data, configurations, vocoders, and even rating scale increments. However, this general decrease in MOS indicates the possibility of range-equalizing bias, and that an unintentional “zooming in” may be taking place as speech synthesis technology improves. It is therefore of utmost importance for speech synthesis researchers to be mindful of their choices of comparison systems and how they may affect MOS results. It is also important when interpreting MOS results to be aware of range-equalizing bias, and to understand that what may appear to be a large difference between systems may in fact be a small difference which has been magnified by a narrow range of quality of synthesized samples in the listening test.

In future work, we will conduct additional listening tests such as A/B and MUSHRA in order to better understand how different tests relate to one another. We will also investigate whether finer-grained listening test data can help MOS predictors to better distinguish between similar systems and to generalize better. We will also investigate zooming in on the lower- and middle-quality regions of MOS, as well as the effect of different rating scales such as ones using increments of 0.5.

6. Acknowledgements

We would like to thank the organizers of the Blizzard Challenges and Voice Conversion Challenges, and the authors of ESPnet-TTS, for making their audio samples freely available. This study is supported by JST CREST Grant Number JP-MJCR18A6 and by MEXT KAKENHI grant 21K11951.

7. References

- [1] "ITU-T Recommendation P.800.2, mean opinion score interpretation and reporting," 2016.
- [2] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.
- [3] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.
- [4] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.
- [5] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.
- [6] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in mushra listening tests," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.
- [7] "ITU-T Recommendation BS.1534-3, method for the subjective assessment of intermediate quality level of audio systems," 2015.
- [8] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 99–104.
- [9] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," in *Proc. 7th International Conference on Speech Prosody 2014*, 2014, pp. 1012–1016. [Online]. Available: <http://dx.doi.org/10.21437/SpeechProsody.2014-191>
- [10] J. O'Mahony, P. O. Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context," in *The 11th ISCA Speech Synthesis Workshop (SSW11)*. International Speech Communication Association, 2021, pp. 148–153.
- [11] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Workshop 2013*, 2013.
- [12] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No!—an empirically-supported critique of Interspeech 2014 TTS evaluations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [13] S. Le Maguer, S. King, and N. Harte, "Back to the future: Extending the Blizzard Challenge 2013," in *Proc. Interspeech*, 2022, pp. 2378–2382.
- [14] S. King, "Measuring a decade of progress in text-to-speech," *Liquens*, vol. 1, no. 1, pp. e006–e006, 2014.
- [15] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice Conversion Challenge 2020 — intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.
- [16] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.
- [17] International Telecommunication Union, Recommendation G.191: Software Tools and Audio Coding Standardization, Nov 11 2005.
- [18] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proc. Interspeech 2017*, 2017, pp. 3976–3980.
- [19] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, and Others, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [20] Y. A. Li, C. Han, and N. Mesgarani, "StyleTTS: a style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.