INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores

Andrew Rosenberg, Bhuvana Ramabhadran

IBM Watson, USA

amrosenb@us.ibm.com, bhuvana@us.ibm.com

Abstract

Listening tests and Mean Opinion Scores (MOS) are the most commonly used techniques for the evaluation of speech synthesis quality and naturalness. These are invaluable in the assessment of subjective qualities of machine generated stimuli. However, there are a number of challenges in understanding the MOS scores that come out of listening tests.

Primarily, we advocate for the use of non-parametric statistical tests in the calculation of statistical significance when comparing listening test results.

Additionally, based on the results of 46 legacy listening tests, we measure the impact of two sources of bias. Bias introduced by individual participants and synthesized text can a dramatic impact on observed MOS scores. For example, we find that on average the mean difference between the highest and lowest scoring rater is over 2 MOS points (on a 5 point scale). From this observation, we caution against using any statistical test without adjusting for this bias, and provide specific non-parametric recommendations.

Index Terms: speech synthesis, listening tests, mean opinion score

1. Introduction

Mean-Opinion score (MOS) listening tests are the most common technique for measuring speech synthesis quality. These listening tests (LTs) are inherently subjective. They ask participants to rate the quality and naturalness of samples of synthesized speech. The subjectivity of the test leads to substantial differences between participants and across different synthesized texts (the utterance). In this work, we demonstrate how large an effect these sources of bias can have, and advocate for the use of non-parametric statistical tests in comparing results from MOS listening tests including specific recommendations for accounting for bias without imposing an incorrect parametric assumption on MOS scores.

In an MOS test, utterances are synthesized from a number of systems. They are presented to a listener and the listener is asked to rate the quality and naturalness of the stimulus on a Likert-type scale from 1 to 5, with scores labeled as Bad, Poor, Fair, Good, Excellent. A range of N of textual utterances are synthesized by each of the M systems, resulting in $N \cdot M$ stimuli in the full test. Depending on how large a test is, a single rater may see a only subset of these. This subset will be approximately (or exactly) balanced between the M systems, but (in general) there is no guarantee that each rater will be presented with the same text synthesized by all candidate systems. A Mean-Opinion Score (MOS) is calculated for each system based on the mean rating of the corresponding responses. The goal of the listening test, and the MOS. is to determine which of the M systems included in the listening test demonstrates the highest performance.

There are some challenges in determining the statistical significance of observed differences between systems. First, the most substantial challenge arises from the fact that LT responses are not strictly "numeric". Rather they are ordinal. This requires the application of non-parametric statistical tests for their analysis (Section 3). Second, there are two known sources of bias in the ratings elicited by an MOS test (Section 4). These stem both from individual differences of raters and differences in the difficulty of specific utterances. Once the sources of bias have been identified there are reasonable non-parametric ways to address these (Section 5). While the observations and recommendations included in this paper are agnostic of any particular listening test, evaluated systems, or synthesized utterances, to make this discussion more concrete, we apply statistical analyses to 46 listening tests that have been run over the last 3 years (Section 6). This allows us to demonstrate the importance of addressing these sources of bias.

2. Related Work

The Blizzard Challenge is the most visible application of listening tests to synthesized speech. Blizzard Challenges have been run every year since 2005 (e.g.[1, 2, 3]). Clark et al. [4] describe a statistical analysis of The 2007 Blizzard Challenge results. In this, they highlight (as we echo) that mean values of Likert-type responses (as are elicited in MOS test) cannot be meaningfully compared to each other. In order to establish a ranking of Challenge participants, the listening tests are structured such that each rater hears examples from each of the M participating systems. This allows the organizers to use Bonferoni-corrected pairwise Wilcoxon signed rank tests to evaluate the differences between systems. This is a non-parametric test, appropriate for paired ordinal data. In such a test, the absolute rating that a listener gives to system A is not important, rather difference between scores given to system A and system B. This pairing allows the analysis to ameliorate the impact of rater bias.

One limitation of this statistical test is that the stimuli from system A and system B will be synthesized from different text utterances (the challenge uses a latin-squares stimulus presentation). This pairing thus exasperates any impact of utterance bias; system A may appear to have higher quality than system B because it is used to synthesize "easier" utterances. To highlight this issue, consider 5 utterances with different inherent difficulties 1, 2, 3, and 4. We synthesize each with systems A and and B, and ask 5 identical raters to listen to one sample from A and one from B. If the raters are presented with the following pairings, (A1, B2), (A2, B3), (A3, B4), (A4, B1), system B would receive higher scores from 3 of 4 raters regardless of the underlying quality of the two systems.

Wester, Valentini-Botinhao and Henter re-examined the result of The 2013 Blizzard challenge in 2015 [5]. While they were not explicitly assessing the question of participant bias,

http://dx.doi.org/10.21437/Interspeech.2017-479

they pointed their analysis on how many subjects are required to obtain a stable ranking of participant systems. They use the same non-parametric paired test as was outlined in [4]. They recommend at least 30 subjects to be used in any listening test – while reporting that most published listening tests base their findings on fewer than 20 participants.

Similar to the contribution here, Ribiero, Florencio, Zhang and Seltzer [6] describe a clear and useful way to adjust variance measure for *numeric* scores for the main sources of bias that can be observed in crowdsourced data collection, namely, participants and stimuli. The aim of this work is to appropriately measure the variance of a MOS statistic, using a two way random effects model with biases from participant preferences, sentence quality (utterances) and general subjective uncertainty. However, this approach is only valid for numeric, and normally distributed, sources of scores and noise, as may be available in a MUSHRA or other test. Since MOS scores are ordinal rather than numeric, this approach is not valid for analyzing this data. A non-parametric approach is required for analyzing MOS tests.

MOS tests are not the only form of listening tests used to assess speech synthesis quality. Other tests more directly compare samples generated from multiple systems. Two common forms of this are ABX and MUSHRA tests. In an ABX test, subjects are presented with two samples, each generated by a different system and are asked to indicate which has higher quality and/or naturalness, sample A, sample B or neither (X). Results are typically reported as a percentage of positive responses obtained by each system. Statistical significance can then be calculated by a proportion test or rank-sum test. MUSHRA tests present stimuli from multiple systems (typically more than 2) and along with a reference stimulus. MUSHRA ratings are elicited on a 0-10 or 0-100 scale. While some versions include adjectival descriptors along this scale, responses are frequently elicited by a slinging bar leading to numeric ratings. Differences here are typically evaluated using a paired t-test. Both MUSHRA and ABX tests present multiple versions of the same utterance to a participant. By pairing samples, the impact of utterance and rater bias is largely mitigated though the use of paired tests.

3. Non-Parametric Statistical Tests

Parametric tests assume a known underlying distribution from which observations (here, participant ratings) are drawn. Typically, as in the case of a t-test, this is a normal or gaussian distribution. There are two problems with using a t-test to analyze MOS results. First, there is insufficient evidence that these scores are drawn from a normal distribution. Second, and more fundamentally, MOS ratings are not *numeric*. Rather, they are *ordinal*, that is, a score of 4 is higher than a score of 3, which is higher than a score of 2. But we have no evidence that the difference between scores of 4 and 3 is the same as the difference between scores of 3 and 2. This renders a test that compares a difference of numeric means to be largely meaningless [7].

Because of this, we suggest the use of the Mann-Whitney U test as a non-parametric test to compare MOS listening test results. This test operates on the ranks of scores rather than their value. It compares the observed difference in the sum of ranks under two conditions, assessing the null hypothesis that there is an equal likelihood that a randomly selected sample from condition 1 will be greater than or less than a sample selected from condition 2. Since ordinal ratings can be ordered (even no assumption of numeric properties), the U test is valid here. Moreover, the U test is very robust to outliers – as the rank of outliers are stable even if the value of the underlying score is large or small.

The Mann-Whitney U test is supported by most (if not all) available statistical packages and toolkits. For MOS listening tests, where it is seldom possible to pair responses, this test is the most appropriate for comparing the performance of evaluated systems. For listening tests, it is only appropriate to pair responses if they have been generated by the same *participant* and contain the same *utterance* (i.e. lexical content). This is due to the bias introduced by these variables (cf. Section 4).

4. Sources of Bias

All subjective tests including listening tests have sources of bias. In this section, we enumerate and describe two of these.

4.1. Participant Bias

Participant bias has been discussed in the context of crowdsourcing extensively in prior work (e.g. [8, 9]). This stems from a healthy skepticism of the fidelity of anonymous participants of uncertain background or expertise that are available in from AMT and other crowdsourcing platforms.

As in all subjective tests, individual participants have biases. These can stem from a number of sources. Some are demographic – age, gender, native language. Some may be derived more from personality – some raters are more forgiving, some are more stringent. Others may be based on a person's experience. A person with a lot of exposure and experience with speech synthesis might expect more from the technology leading to a more judgmental rater, while someone with less experience with this technology may be more forgiving. They also may have different degrees of interest in the task. Other factors than can be revealed as participant bias are more environmental. Raters (particularly when solicited via crowdsourcing platforms) use different listening equipment and they take the test in different environments. Interference from listening equipment and environmental noise can impact assessments of stimuli, especially of "quality".

4.2. Utterance Bias

All synthesis technology is better at synthesizing some utterances than others. In the extreme, if an utterance appears in the training material, the synthesis quality will be maximally good. This isn't an unreasonable possibility; many common short phrases are intentionally included in source material for this reason. Qualities of lexical content can pose challenges to many points along the synthesis pipeline. From the text-analytics (e.g. syntactic analysis, and text normalization), prosodic assignment and pronunciation generation to the backend audio generation process whether parametric or concatenative.

We define "utterance bias" here as the quality that based solely on the lexical content being synthesized, some utterances will be scored more highly than others. It could be argued that utterance variation is desirable in a listening test, and certainly, multiple and varied utterances are used in any evaluation to more thoroughly probe the differences between systems. If all utterances are presented under all systems an equal number of times, any impact of utterance bias on listening test results will be minimal – all systems will benefit or suffer equally. However, it is quite common to randomly assign a subset of system/utterance pairs to a user in order to reduce their workload and/or evaluate more utterances or systems than one listener could reasonably be expected to judge. In this case, utterance bias can impact system scores.

5. Addressing Bias

Just as MOS tests require non-parametric tests for determining statistically significant differences, a non-parametric approach to addressing bias is necessary. In this section, we describe normalized-rank normalization, a simple non-parametric method to address these sources of bias in analyzing the results of MOS listening tests. This normalization approach is well motivated for ordinal data, and for numeric data whose generating distribution is not known.

Rank normalization is a technique a the center of the Mann-Whitney U test. Under this scheme scores are converted to their "rank" if they were sorted. This is accomplished by assigning numeric ranks to all observations, starting at 1 for the smallest, up to N for the largest. When there are ties – multiple observation of the same score – all receive the same rank, that of the midpoint of the range of ranks they would cover. For example scores [1, 2, 2, 2, 4, 5, 5] become ranks [1, 3, 3, 3, 5, 6.5, 6.5]. The 2's range cover ranks 2 through 4, yielding a midpoint of 3. The 5s cover ranks 6 and 7, giving a midpoint of 6.5. Normalized-rank normalization ensures that ranks vary from 0 to 1, rather than 1 to N, where N is the number of observations. To normalize ranks, one subtracts 1 and divides by N-1. Our example scores then become [0/6, 2/6, 2/6, 2/6, 4/6, 5.5/6, 5.5/6]. While this normalization approach is not novel, we believe its application to this task is.

It is necessary in the data used here (cf. Section 6) to use the normalized-rank version because some participants perform more than one HIT leading them to make more observations. If a listening test includes the same number of ratings per participant, vanilla rank normalization can be used.

To calculate statistical significance with participant bias correction, we first apply normalized-rank normalization to each participant's scores separately. We use a Mann-Whitney U test to compare the scores applied to each system. We use the same approach for utterance bias correction; we normalized-rank normalize each utterance's scores, then use a Mann-Whitney test. To correct for both sources of bias, we first normalize by speaker, then perform a second rank normalization by utterance. We proceed in this order because we find speaker bias to be greater than utterance bias (Section 7) but we don't see a large difference if the order is reversed.

6. Material

To ground the previous material in real results, we have identified 46 legacy listening tests. All of these were run on Amazon Mechanical Turk (AMT) between October 2015 and February 2017. All participant scores are in response to a prompt to "rate the overall quality and naturalness" of a given stimuli. The 5-point scale labels were 1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, and 5 - Excellent for all tests. The number of evaluated systems varies between 2 and 7 (though only 10 tests have 4 or more). We only include the first two systems in each test in our analysis. If we were to use all pairs of systems for comparison, larger listening tests would have outsized influence on the following discussion, as there are $\binom{N}{2}$ combinations of N systems.

Six out of the 46 investigated listening tests involved voices in a language other than US English these included two Spanish tests, two French tests, one German and one UK English. Since AMT is an English language site, we assume that participant's have a working knowledge of English, and present the instructions identically in the US English and non-US English tests. On non-US English Listening tests, we only accept participants who self-report to be native speakers of the stimuli language.

In addition to quite restrictive participant requirements in terms of number of previously completed tasks and their acceptance rate, we also perform outlier removal on participants of each listening test. The outlier removal process we use is based on the (linear) correlation of rater scores, and proceeds iteratively. We start with a list of "coherent" raters C and an empty list of outliers O. For each rater, $r \in C$, we calculate the pearson correlation coefficient ρ_r between ratings from r and the mean ratings of $(C \setminus r)$ on corresponding stimuli. Then again for each rater, r, we then construct hypothetical C' and O' sets where $C' = (C \setminus r)$ and $O' = O \cup r$. We calculate a "coherency gap" measure, g, for each C' and O' partition where

$$g = \frac{\min_{r \in C'} \rho_r - \max_{r \in O'} \rho_r}{(\max_{r \in C'} \rho_r - \min_{r \in C'} \rho_r)/|C'|}.$$

This gap is the distance between the rater in C' with the lowest agreement with the group and the rater in O' with the highest agreement. The size of this gap is normalized by the range of the correlations in C' normalized by the size of C'. At each iteration, n, we move the rater r that generated the largest g from C to O. We call this gap g_n. We repeat this process until either more than 15% of subjects are moved to O, or the lowest correlation among $r \in C$ is greater than 0.45 times the maximum correlation observed during the first iteration. At termination, we identify the iteration n' that resulted in the largest gap g'_n, and identify the corresponding outlier set O at iteration n as the final set of outliers.

7. Measuring Bias

In any subjective test, there will inevitably be bias that can have the effect of corrupting participant ratings. Here, we seek to measure the amount of bias we can observe as a function of participant identity and utterance content. We will also report on observed priming effects.

To measure participant bias, we calculate the mean participant score μ_p, for each listening test, $p \in P$. We describe the variance of participant behavior in two ways: 1) we calculate the standard deviation of these mean scores per listening test, $\sigma(\{\mu_p \forall p \in P\})$ and 2) we calculate the spread, ρ_p, as the difference between the highest and lowest *mean* rating $\rho_p = \max_p(\mu_p) - \min_p(\mu_p)$. We then inspect the mean spread across all tests, $\mu(\{\rho_p \forall p \in P\})$

Averaged across all 46 tests, we find that the average standard deviation of participant mean scores to be 0.484. This indicates a very high level of subject bias. By considering the mean participant score, we have already eliminated a considerable amount of variance. If we assume that 2/3 of raters fall within 1 standard deviation of this mean, this indicates 1/3 of all subjects have a mean score that is almost 0.5 higher or lower than the reported MOS. Considering (merely anecdotally) that an MOS difference of 0.2 is considered "large", and fairly consistently calculated to be statistically significant, the impact of this variance is quite substantial. When comparing the range of rater means, we find that on average the spread is 2.31 with the smallest observed spread being 1.7 and largest being 3.2. Considering that the MOS scale ranges from 1 to 5, an expected difference of *mean* as large as 2.31 is enormous. In Figure 1 we include a visual representation of participant responses drawn from the listening test with the median participant spread, 2.3. This figure includes the mean and standard deviation for each participant along with the overall population mean and standard deviation in red. Figure 2 shows the corresponding normalized-rank adjusted rater statistics.

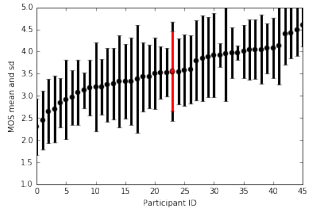

Figure 1: *Example sorted MOS per participant. Overall mean (3.56) and s.d. (0.90) appear in red.*

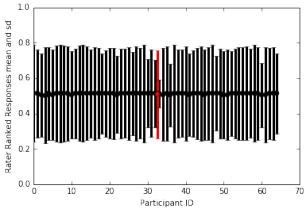

Figure 2: *Norm.-rank normalized MOS per participant.*

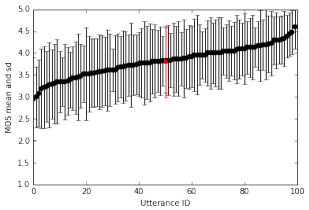

Figure 3: *Example sorted MOS per utterance. Overall mean (3.81) and s.d. (0.82) appear in red.*

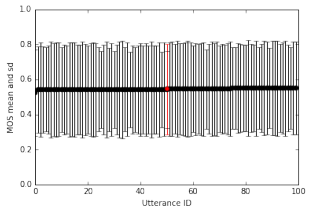

Figure 4: *Norm.-rank normalized MOS per utterance.*

Recall that six listening tests involved languages other than US English. The participant bias measures of these are consistent with the overall findings – an average spread of 2.35 and an average standard deviation of means of 0.497. It is worth noting that the largest observed spread of 3.2 does come from the UK English test, though there are US English tests with spreads of 3.075 and 3.0. Because of this, we don't believe the UK English result to be a function of the language of the examined voice.

That participant bias exists shouldn't be surprising, though the magnitude is remarkable. To see if this was unique to crowdsourcing and the AMT platform, we identified two, rather large (71 and 64 participants) traditional listening tests. These were performed by IBM employees, in their own offices. We find similar participant biases on these tests: participant mean spreads of 2.75 and 3.0 and standard deviations of 0.636 and 0.509. This suggests that participant bias effects may not be specific to AMT, but may be endemic in MOS tests.

We now turn our attention to utterance bias. We measure utterance bias in the same way as we previously measured participant bias. We calculate the mean rating for each utterance. From this we measure the "spread" as the difference between the maximum and minimum mean utterance rating in a listening test. We also measure the standard deviation of utterance means within each test. Here we find that across all 46 tests, the average spread of utterance scores is 1.47, with a standard deviation of 0.33. While participant bias is frequently a source of concern in so-called "crowdsourced" applications of listening tests, we find that the impact of the lexical content to be substantial nearly to the extent of participant bias. Figure 3 includes a chart of mean utterance scores drawn from the listening test with the median spread. The mean and standard deviation of scores for each utterance and the test mean and standard deviation are included in red. Figure 4 includes the corresponding utterance scores following normalized-rank adjustment. Repeating this on the two internal tests reveals variance slightly higher than the average observed on the AMT tests – with spreads of 1.71 and 2.02 and standard deviations of 0.369 and 0.426.

One reason utterance bias is often not considered to be as serious a problem as participant bias is that it is frequently controlled for across systems by including examples from all systems synthesizing each utterance. While this may have a limited impact in measures of the statistical significance of differences between systems drawn from the same test, it draws a clear light on why MOS results from one set of utterances should be compared to results from a different set of utterances with *extreme* prejudice. This also draws light on the fact (well-known anecdotally, but perhaps under-reported) that the make up of listening test stimuli can have a sizable impact on MOS scores.

8. Impact on Listening Tests

Here we describe how the proposed bias-correcting methods (Section 5) impact statistical significance of listening test results. Using the set of 46 listening tests described in Section 6, a vanilla Mann-Whitney U test shows that 24 tests have statistically significant ($p \leq 0.01$) differences between investigated systems. When we adjust for participant bias, the same 24 tests are statistically significant. However, the test statistic, U, increases on 42 out of 46 tests. While we are not observing cases where the impact of rater bias is leading to erroneous judgements of statistical significance, we can see clear evidence that when we eliminate this bias systems are judged to be more different than if we ignore this source of bias.

We then adjust for utterance bias. When we do this, the same 24 tests are considered statistically significant at the $p \leq 0.01$ level. Here 36 of 46 tests show an increased test statistic following utterance bias adjustment.

Finally, we adjust for *both* participant and utterance bias. Here we see two additional tests being statistically significant, and 41 out of 46 tests have an increased test statistic. This suggests observed MOS scores and measures of statistical significance of the distance between systems are measurably impacted by both participant and utterance bias.

Note, we haven't compared the results from the Mann-Whitney U tests to (parametric) Student's t-tests. Despite is being occasionally used to compare MOS scores, the t-test is not appropriate to analyze this kind of data (cf. Section 3). We believe including it here, despite its clear flaws, would serve to continue its erroneous justification in the analysis of MOS tests.

9. Conclusions

This paper highlights the impact of two common sources of bias in MOS listening tests: participant bias and utterance bias. We describe these sources of bias and measure their presence in a set of 46 legacy listening tests. We find there to be an incredibly large variance in mean participant scores and mean utterance scores. This suggests that statistical tests comparing the differences between listening tests should be used to account for this bias. Knowing these sources of bias are pervasive in subjective tests, we advocate for the use of non-parametric techniques to mitigate bias and appropriately assess the significance of observed differences and demonstrate their impact on the set of investigated tests. We find in 41 of 46 tests, bias adjustments increase the statistical significance of observed differences suggesting that these are reliable sources of noise.

10. References

[1] A. Black, "The blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *SSW6*, 2005.

[2] S. King and V. Karaiskos, "The blizzard challenge 2013," in *SSW8*, 2013.

[3] ——, "The blizzard challenge 2016," in *SSW9*, 2016.

[4] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," in *in Proc. Blizzard Workshop (in Proc. SSW6*, 2007.

[5] M. Wester, C. Valentini-Botinhao, and G. Henter, *Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations*. International Speech Communication Association, 9 2015, pp. 3476–3480.

[6] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: an approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2416–2419.

[7] H. M. Marcus-Roberts and F. S. Roberts, "Meaningless statistics," *Journal of Educational Statistics*, vol. 12, no. 4, pp. 383–394, 1987.

[8] F. L. Wauthier and M. I. Jordan, "Bayesian bias mitigation for crowdsourcing," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 1800–1808. [Online]. Available: http://papers.nips.cc/paper/4311-bayesian-bias-mitigation-for-crowdsourcing.pdf

[9] E. Kamar, A. Kapoor, and E. Horvitz, "Identifying and accounting for task-dependent bias in crowdsourcing," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.