RELIABLE ESTIMATION OF QUALITY SCORES BY A SMALL CALIBRATED LISTENING PANEL

Iman Mossavat

Eindhoven University of Technology
Department of Electrical Engineering
Eindhoven, The Netherlands

W. Bastiaan Kleijn

Victoria University of Wellington
School of Engineering and Computer Science
Wellington, New Zealand

ABSTRACT

In this work we propose the calibrated-mean-opinion-score (CMOS), which accounts for varying levels of precision and bias across subjects in a listening test. We adopt the Bayesian statistical framework, where the hyper-parameters of priors are learned via empirical Bayes, and the posterior is approximated by a computationally inexpensive variational technique. As our experimental results show, CMOS is more robust to noisy and biased subjects than MOS. As a result, CMOS can be used to improve the reliability of listening test results when a small test panel is used. To correct for the subjects in the test panel, calibration signals are required. Calibration signals are rated by a panel larger than the test panel. The key to saving human labor and cost is that only a few calibration signals are required, and that it is possible to share calibration signals across listening tests.

Index Terms— Listening test, Speech quality, Bayesian inference, bias

1. INTRODUCTION

Perceived quality of the speech or audio signal is often a key performance indicator for communication network designers and DSP engineers. Subjective quality tests are the most accurate way to assess the quality of audio and speech [1, 2]. The main disadvantages of subjective tests are cost, human labor, and time. Objective quality measures provide an estimate for the speech or audio quality, but with less accuracy in comparison with subjective listening tests [3–7].

Subjective ratings are influenced by a number of factors including hearing, perception, judgment and mapping (translation of internal judgment into a quality score) [2]. All of the aforementioned factors potentially introduce biases, i.e. systematic shifts in the scores. Personal expectations, emotions and mood of the subject affect the scores and produce bias [8] [9] [10]. Mapping of judgments into scores introduces bias in several ways including but not limited to contraction [11], centering [12] [13], range equalizing [14], and stimulus spacing [15] [16] biases.

Reducing the biases is in general a difficult task. Biases such as centering and range equalizing are considered as the most difficult type to reduce [17]. One way to reduce centering and range equalizing biases is monadic (single stimulus) tests such as Mean-Opinion-Score (MOS) [18], which generally require large panel size to gain statistical significance [17]. Systextual design [19] systematically manipulates the range and distribution of stimuli, and is another approach to reduce centering and range equalizing biases at the cost of increasing the number of stimuli. Contraction bias is reduced by multiple stimulus tests such as recommendation ITU-R BS.1534-1 (MUSHRA) [20] or by a technique called direct anchoring [21]. MUSHRA is known to be vulnerable to range equalizing bias [22].

In this paper we propose a method to reduce the bias of tests for a relatively small panel size. The main idea behind the method is that it is not necessary for the large panel to rate each and every test signal. Instead, it is sufficient to use the large panel size only for rating a 'calibration set': a limited number of signals (around 10 signals). We name the large panel the 'calibration panel'. It is, in principle, possible to share the calibration set across several tests. The 'test panel' is the panel that rates the quality of the test material as well as the calibration set. In our approach, the noisiness and systematic shifts of test panel subjects are characterized by jointly modeling the quality ratings of the test panel and the calibration panel over the calibration set. The statistical characterization is then used to correct for the expected errors in the test panel ratings. Combination of our method with single stimuli testing such as [18] allows for reduction of centering and range equalizing bias, while limiting the test panel size. Also, biases due to affective judgments (expectations, mood, and emotions) are reduced by our method since the responses are corrected for. We use variational Bayesian statistical inference methods to characterize the subjects.

Bayesian statistics facilitates the transfer of knowledge from one listening experiment to the other via Bayesian priors. For example, we successfully characterize the subjects in one database by using the prior context generated using other databases of Supplement 23. Mathematically, priors are es-

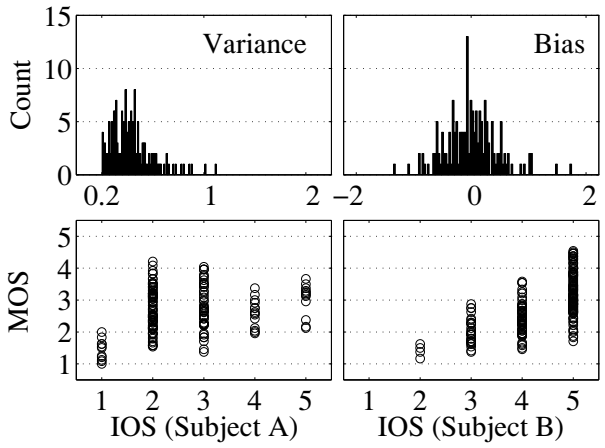

Fig. 1. Top Left: Variance of error (error = IOS-MOS) Top Right: Bias (mean value) of error Bottom Left: IOS versus MOS for a noisy subject Bottom Right: IOS versus MOS for a biased subject

sential to shrinking the size of the calibration set, which is the key to reducing the expensive human labor. Our work not only motivates the addition of calibration sets to existing standardized procedures, but also motivates standardized datasets that can provide 'calibration' signals and calibrated ratings.

The organization of the paper is as follows: in Section 2 we introduce the model, and in Section 3 the variational Bayesian inference algorithm is presented. In Section 4 we present the experimental results using Supplement 23 and NOIZEUS. In Section 5 we discuss and analyze the implications of the study and conclude the paper.

2. MODEL

The widely used MOS is the average of Individual-Opinion-Scores (IOSs):

$$\mathrm{MOS}_s = \frac{1}{m}\sum_{i=1}^{m} \mathrm{IOS}_{is}$$

where IOS_{is} is the rating given by individual $i = 1, \cdots, m$ to speech sample $s = 1, \cdots, N$. We assume that for each sample there is a true-score (TS), and that the following model relates the IOS to TS:

$$\begin{aligned} \mathrm{IOS}_{is} &= \mathrm{TS}_s + n_{is} + b_i \\ \mathrm{Prob}(n_{is}) &= \mathcal{N}(n_{is}|0, \lambda_i^{-1}) \end{aligned} \tag{1}$$

where n_{is} is the i-th subject Gaussian noise with variance λ_i^{-1}, and bias b_i. In practice the relation between IOS_{is} and TS_s is more complex as subjects use a discrete finite scale of one to five to express their opinion and their ratings are not Gaussian distributed. However, we will show that our model results in improved accuracy (with respect to MOS) in the estimation of the underlying latent score TS_s from observable rates IOS_{is}. CMOS is the estimation of TS based on the data and priors. Equation (1) implies that the precision λ_i and the bias b_i of subjects are stationary across signals.

Histograms of bias and variance of Supplement 23 subjects are plotted in upper panels of Figure 1. Histograms illustrate that variance and bias change across Supplement 23 subjects. The scatter plots of IOS versus MOS for the most imprecise subject (A) and the most biased subject (B) are shown in lower panels of Figure 1. IOS of subject A indicated very little information about the MOS. Subject B does not rate any speech signal as 'bad', and IOS of subject B is always lower than MOS. As we will show later, Bayesian inference adjusts the weight of subjects in CMOS based on their noise level, and corrects for the shifts in the scores of each subject.

3. STATISTICAL INFERENCE FRAMEWORK

Compared to the standard MOS framework [18], our model introduces additional complexity in the form of bias and variance variables for each subject. To maintain the generalization capability of our model, we use a Bayesian formalism, which uses priors to regularize the inference problem. Our choice of priors allows for an efficient variational estimation of the posterior. In other words, we avoid sampling of the posterior by deliberately choosing priors with suitable mathematical properties.

The Gaussian-Gamma prior over bias and precision is defined to be

$$\mathrm{Prob}(b_i, \lambda_i|\beta) \quad = \quad \mathcal{N}\left(b_i|0, (\beta\lambda_i)^{-1}\right) \mathrm{Gamma}\left(\lambda_i|a_0^{\lambda}, b_0^{\lambda}\right) \tag{2}$$

where

$$\mathrm{Prob}(\beta) \quad = \quad \mathrm{Gamma}\left(\beta|a_0^{\beta}, b_0^{\beta}\right) \tag{3}$$

where a_0^{λ}, b_0^{λ}, a_0^{β} and b_0^{β} are 'hyper-parameters' set before training that specify the strength of the priors. The hyper-parameters roughly determine how noisy a subject can be. We will discuss how hyper-parameters are set before training. Variable β does not correspond to an actual physical concept and its role is to add flexibility to the joint prior over b_i and λ_i [23].

The priors in Equations (2) and (3) are conjugate to the likelihood function derived from the model in Equation (1), which allow us to derive an efficient iterative variational inference algorithm for estimating the posterior distribution over parameters. The variational method is known as the mean-field posterior approximation [24] [25] [23]. The derivation is standard and we will not present it. We only state the iterative equations as they provide insight into the implications of

model in Equation (1),

$$\begin{aligned} \mu_s &= V_s \sum_{i=1}^{m} \widehat{\lambda}_i \left(\mathrm{IOS}_{is} - \widehat{b}_i \right) \\ V_s^{-1} &= \sum_{i=1}^{m} \widehat{\lambda}_i \end{aligned} \quad (4)$$

where μ_s is the estimate of the model for TS_s which is updated iteratively based on estimates of bias $\widehat{b}_i$ and precision $\widehat{\lambda}_i$ for all subjects. Note how Equation (4) differs from averaging done for MOS: the estimates of individual biases are subtracted from the ratings, and each rating is weighted by the precision of the subject. More precise subjects have heavier weight in the estimate.

The iterative estimates of bias and precision are given by

$$\begin{aligned} \widehat{b}_i &= V_b \sum_{s=1}^{N} (\mathrm{IOS}_{is} - \mu_s) \\ V_b^{-1} &= (N + \widehat{\beta}) \end{aligned} \quad (5)$$

where $\widehat{\beta}$ is the estimate of β at the last iteration. In Equation (5) the bias estimate for a subject is proportional to the discrepancy between the subject ratings and the estimates of true quality averaged over all samples. Since over-fitting is a concern as more parameters are introduces, it is desirable that the degree of freedom in Equation (1) provided by the bias term is used only when sufficient data is available. Asymptotically, as number of speech samples, N increases, the value of $\widehat{\beta}$ become less relevant. On the other hand, the bias estimates shrink toward zero because of $\widehat{\beta}$ for small N.

The estimates of subject precision is given by

$$\begin{aligned} \widehat{\lambda}_i &= a^{\lambda}/b^{\lambda} \\ a^{\lambda} &= a_0^{\lambda} + N/2 \\ b^{\lambda} &= b_0^{\lambda} + 0.5 \sum_{s=1}^{N} \mathbb{E}\left\{ (\mathrm{IOS}_{is} - \mu_s)^2 \right\} \\ &\quad -0.5 V_b \left\{ \sum_{s=1}^{N} (\mathrm{IOS}_{is} - \mu_s) \right\}^2 \end{aligned} \quad (6)$$

As Equation (6) shows, the precision estimate depends on the second order moments of differences between subject ratings and the estimates. Note that for small N, the hyper-parameters a_0^{λ} and b_0^{λ} reduce the dependency of precision estimates to data.

Finally the updates for iterative estimates of β are given by:

$$\begin{aligned} \widehat{\beta} &= a^{\beta}/b^{\beta} \\ a^{\beta} &= a_0^{\beta} + m/2 \\ b^{\beta} &= b_0^{\beta} + 0.5 m V_b + 0.5 \sum_{i=1}^{m} \widehat{\lambda}_i \widehat{b}_i^{\,2} \end{aligned} \quad (7)$$

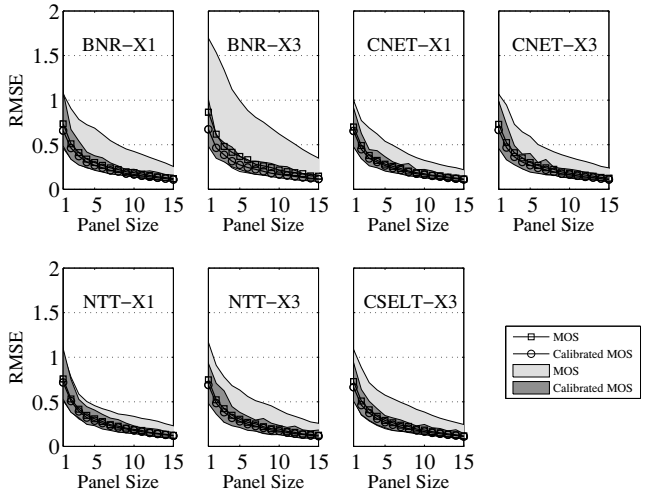

Fig. 2. The curves show the average performance of MOS and CMOS over 100 panels. The smallest- and largest- RMSE determine the boundaries of the shaded areas.

The algorithm iteratively updates the estimates of model parameters in Equations (4), (5), (6), and (7) until convergence is achieved.

4. EXPERIMENTAL RESULTS

In this Section we demonstrate the effectiveness of CMOS. We show how CMOS prevents large errors when small subsets are used, and we present a brief study of calibration signals and hyper-parameters.

4.1. Dataset

ITU-T Supplement 23 is one of the few publicly available datasets. We test our method on experiments one and three of Supplement 23, which contain absolute category rating (ACR) tests. In the ACR test subjects grade the speech samples on a discrete opinion scale: 'Excellent', 'Good','Fair','Poor','Bad'. The data consist of seven databases with speech samples from different languages distorted by different conditions. The panel size is 24 subjects.

4.2. RMSE Performance

In our first experiment, we considered test panel sizes of 1 to 15. For each of the seven databases of Supplement 23 and for each panel size $m > 1$, we constructed 100 test panels by randomly drawing m subjects from the pool of 24 subjects in that database. We defined the error as the difference of the results obtained from all 24 subjects and the results obtained from the test panel. We randomly drew 10 signals from the database to form our calibration set. We used all 24 subjects as the calibration panel. Prior to testing each database, we used the

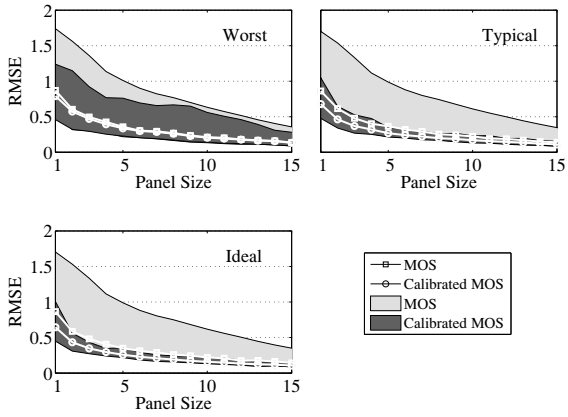

Fig. 3. Effect of calibration set on RMSE. The area between the smallest and largest RMSE is shaded.

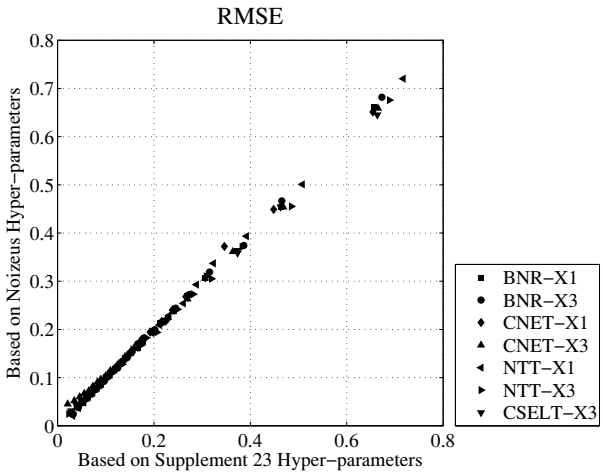

Fig. 4. Effect of hyper-parameters on RMSE

remaining databases to learn the hyper-parameters a_0^λ, b_0^λ, a_0^β and b_0^β by fitting the model to the databases, and deploying the empirical Bayes method [26].

We limited the panel size to 15 for two reasons: firstly, to demonstrate the value of our method the calibration panel must be larger than the test panel. Secondly, as the number of subjects increases, the MOS of test panel reaches the MOS of 24 subjects, and the error is not meaningful for comparing the generalization performance of our method against MOS.

Figure 2 illustrates the performance of MOS as well as our method, i.e. the calibrated MOS (CMOS). The MOS and CMOS curves (averaged over 100 random panels) are plotted, and the areas between the smallest- and largest-RMSE values are shaded. CMOS slightly, but consistently performs better than MOS in terms of mean. The main advantage of CMOS is lowering the largest RMSE.

4.3. Calibration Set

To demonstrate the role of calibration set we considered three calibration sets: 1 - To generate the 'worst case' calibration scenario, we chose the 10 signals with the smallest value of $\sum_{i=1}^{m} b_i.(\mathrm{IOS}_{is} - \mathrm{MOS}_s)$, where m is the number of subjects in calibration panel, and s denotes the index of the signal. All database subjects and signals were used to calculate the bias b_i. Thus, the signals were chosen to generate a large error in bias estimation. 2 - In the 'ideal case' scenario we used all the data (all subjects and all signals) to characterize the subjects, which resulted in the most accurate estimation of bias and variance given the data. 3- In the 'typical case' we randomly drew 10 signals as our calibration set.

In Figure 3 the effect of the calibration set on RMSE is illustrated for database BNR-X3. The first major observation is that in the worst case CMOS still compares favorably against the MOS. The second important observation is that 10 randomly drawn signals perform quite closely to the ideal case, which indicates that the choice of calibration signals in this experiment is not critical.

	Supplement 23	NOIZEUS
a_0^λ	7.30	9.36
b_0^λ	2.89	3.75
a_0^β	5.75e-005	3.57e-005
b_0^β	0.012	0.011

Table 1. Hyper-parameters learned from Supplement 23 and NOIZEUS.

4.4. Effect of hyper-parameters

To demonstrate the potential of priors in transferring information from one dataset to the other, we did an experiment where we used the data in one dataset with noisy speech samples (NOIZEUS) to generate hyper-parameters that are used to characterize the subjects in another dataset, Supplement 23. In Table 1 we see the values of hyper-parameters learned from Supplement 23 and NOIZEUS are similar. Figure 4 shows that similar RMSE performance is achieved regardless of the data used to learn the hyper-parameters.

5. CONCLUSION

We presented a method to calibrate subjects and their MOS. The decrease of maximum error shows that calibration improves the robustness of the results against noisy and biased subjects, which in turn allows for smaller panel sizes. Under the testing conditions in Supplement 23, only 10 randomly drawn calibration signals offer a significant drop of the largest errors. We showed that existing listening data is useful in constructing priors to calibrate subjects in new tests.

6. REFERENCES

[1] Volodya Grancharov and W. Bastiaan Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*, J. Benesty, A. Huang, and M. Sondhi, Eds., chapter 5, pp. 83–102. Springer, Nov. 2007.

[2] S. Bech and N. Zacharov, *Perceptual Audio Evaluation-Theory, Method and Application*, Wiley, 2007.

[3] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, NJ:Prentice-Hall, 1988.

[4] J.G. Beerends, A.P. Hekstra, A.W. Rix, and M.P. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model," *J. of the Audio Eng. Society*, vol. 50, no. 10, pp. 765–778, 2002.

[5] L. Malfait, J. Berger, and M. Kastner, "P. 563-the ITU-T standard for single-ended speech quality assessment," in *IEEE Trans. on Audio, Speech and Lang. Process.*, 2006, vol. 14, pp. 1924–1934.

[6] I. Mossavat, P.N. Petkov, W.B. Kleijn, and O. Amft, "A hierarchical Bayesian approach to modeling heterogeneity in speech quality assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 136–146, 2012.

[7] S. I. Mossavat, O. Amft, B. de Vries, P. N. Petkov, and W. B. Kleijn, "A Bayesian hierarchical mixture of experts approach to estimate speech quality," in *Int. Workshop on Quality of Multimedia Experience*, 2010.

[8] D. Västfjäll and M. Kleiner, "Emotion in product sound design," *Proceedings of Journées Design Sonore*, 2002.

[9] D. Västfjäll, "Contextual influences on sound quality evaluation," *Acta acustica united with acustica*, vol. 90, no. 6, pp. 1029–1036, 2004.

[10] K. Zimmer, W. Ellermeier, and C. Schmid, "Using probabilistic choice models to investigate auditory unpleasantness," *Acta acustica united with acustica*, vol. 90, no. 6, pp. 1019–1028, 2004.

[11] K. Beresford, N. Ford, F. Rumsey, and S.K. Zieliński, "Contextual effects on sound quality judgements: listening room and automotive environments," in *Journal Audio Engineering Society (Abstracts)*, 2006, vol. 54, p. 666.

[12] F.E. Toole, "Listening tests–turning opinion into fact," in *Journal Audio Engineering Society (Engineering Reports)*, 1982, vol. 30, pp. 431–445.

[13] H. Helson, *Adaptation-level theory*, Harper & Row, 1964.

[14] H.T. Lawless and H. Heymann, *Sensory evaluation of food: principles and practices*, Kluwer-Plenum, 1998.

[15] B.A. Mellers and M.H. Birnbaum, "Loci of contextual effects in judgment.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, no. 4, pp. 582–601, 1982.

[16] A. Parducci, *Happiness, pleasure, and judgment: The contextual theory and its applications.*, Lawrence Erlbaum Associates, Inc, 1995.

[17] S. Zieliński, F. Rumsey, and S. Bech, "On some biases encountered in modem audio quality listening tests: A review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[18] "Mean opinion score (MOS) terminology," ITU, Geneva, Switzerland, 2003, ITU-T Rec. P.800.1.

[19] M.H. Birnbaum, "Controversies in psychological measurement," *Social attitudes and psychophysical measurement*, pp. 401–485, 1982.

[20] "ITU-R BS.1534-1 Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunication Union, Geneva, Switzerland, 2003.

[21] J.P. Guilford, *Psychometric methods*, McGraw-Hill, 1954.

[22] S. Zieliński, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," 2007, convention paper 7179.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[24] T.S. Jaakkola and M.I. Jordan, "Bayesian parameter estimation via variational methods," *Stat. and Computing*, vol. 10, no. 1, pp. 25–37, 2000.

[25] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Univeristy Press, 2003.

[26] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, Texts in Statistical Science. Chapman and Hall, First CRC Press reprint 2000, 1995.