INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Analysis of mean opinion scores in subjective evaluation of synthetic speech based on tail probabilities

Yusuke Yasuda[1], *Tomoki Toda*[1]

[1]Nagoya University, Japan

yasuda.yusuke@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Subjective evaluations such as mean opinion scores (MOS) are essential for evaluations of synthetic speech including automatic speech quality assessment (SQA) models. In this paper, we evaluate the confidence intervals of MOS in a listening test and the number of required samples to achieve a certain confidence interval based on various tail probability evaluation methods. The tail probability is a probability representing the sample mean deviates greatly from the true mean. We use tail probability evaluations based on asymptotic and upper-bound-based approaches. In our experiments about toy data and actual listening test data, we show that achieving small confidence intervals requires huge sample volumes, and the MOS corpus for SQA has large confidence intervals due to limited sample volumes. We suggest adopting comparative scoring and online learning for more reliable subjective evaluations under limited budgets as the future direction.

Index Terms: subjective evaluation, tail probability, confidence interval, mean opinion score, MOS

1. Introduction

The quality of synthetic speech has been improved significantly by the advancement of deep learning. This improvement has reached the level where some methods achieve the naturalness of synthetic speech close to that of natural speech [1, 2, 3]. It is expected that this trend of quality saturation of synthetic speech is making subjective evaluations hard to distinguish qualities with sufficient probability under limited samples [4]. Recently, corpora collecting mean opinion scores (MOS) are proposed to realize model-based automatic speech quality assessment [5, 6]. These MOS corpora can be used as training data for MOS predictors, and many MOS prediction methods are proposed [7, 8, 9, 10]. Analysis of listening test results is therefore increasing its importance because of two factors: (1) the limit of reliability of the current listening test scheme may be approaching due to the quality saturation and (2) the reliability of the MOS corpora limits the performance of the automatic quality evaluation models.

In this paper, we evaluate the confidence intervals of MOS in a listening test and the number of required samples to achieve a certain confidence interval based on various tail probability evaluation methods. The tail probability is a probability representing the sample mean deviates greatly from the true mean. We describe tail probability evaluations based on an asymptotic approach in Section 2.1 and the upper-bound-based approach in Section 2.2 and summarize them in Section 2.3. In Section 3, we apply the tail probability evaluation methods to toy data and actual listening test data. Section 4 concludes our finding and provides the future direction of subjective evaluation. Our contributions are as follows:

- We show true confidence intervals are expected to be larger than common confidence intervals based on the central limit theorem due to underestimation;
- We show that achieving small confidence intervals requires huge sample volumes that are not collectible with crowdsourcing;
- We show that MOS from an existing listening test corpus has large confidence intervals that indicate systems with similar MOS can not be ranked with sufficient probability.

2. Tail probability evaluation

In evaluations of synthetic speech, we estimate a true expectation of the quality of the synthetic speech system with a sample mean in the form of MOS. The sample mean always contains estimation errors because we estimate the unknown true mean with limited samples. With a small probability, the sample mean becomes far from the true mean. We can evaluate a probability where the sample mean deviates greatly from the true mean, and the probability is called tail probability. We usually report the degree of the estimation errors of MOS with confidence intervals based on the tail probability. In the following sections, we describe asymptotic and upper-bound-based tail probability evaluations.

2.1. Asymptotic approach

2.1.1. Central limit theorem

The central limit theorem is one of the major tail probability evaluation methods. It is commonly used to evaluate the confidence interval of MOS.

In the central limit theorem, the normalized sample mean is converged in normal distribution at the limit of the number of samples n. Therefore, tail probability can be evaluated with cumulative normal distribution:

$$\lim \mathbb{P}_{n\to\infty}[\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \leq x] = \Phi(x) \tag{1}$$

where $\hat{\mu}_n$ is sample mean, μ is true mean, σ is true standard deviation, and $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ is cumulative normal distribution.

Confidence interval Δ can be evaluated as follows:

$$\Delta = \hat{\mu}_n - \mu = \Phi^{-1}(\delta/2)\frac{\sigma}{\sqrt{n}}, \tag{2}$$

where δ is the error probability. One of the common values of the error probability is 0.05, which results in a 95% confidence interval.

The number of samples to achieve error probability δ can be derived as follows:

$$n = \left[\Phi^{-1}(\delta/2)\frac{\sigma}{\Delta}\right]^2 \tag{3}$$

10.21437/Interspeech.2023-1285

The evaluation of the confidence interval with normal distribution requires the true standard deviation to be known. Thus, Student's t-distribution is usually used to calculate the confidence interval instead of a normal distribution to replace the true standard deviation with the sample standard deviation.

$$\Phi'(x) = \int_{-\infty}^{x} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} dt \quad (4)$$

Here, Γ is the Gamma function, and ν is a degree of freedom. The Student's t-distribution can represent a distribution of the sample mean when $\nu = n - 1$.

A problem of the central limit theorem is the low accuracy of tail probability approximation when n is small. Berry–Esseen theorem [11] indicates that there is an approximation error proportional to $1/\sqrt{n}$ in the central limit theorem. In concrete, let x be a random variable with 0 means, σ standard deviation, and ρ third-order absolute moment, and let $F_n(x)$ be a cumulative distribution of normalized sample mean of x. Then, the normal approximation error is bounded by the following formula:

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}, \quad (5)$$

where C is a positive constant. The approximation error of the central limit theorem could be large when n is small because it gets smaller only as fast as $O(1/\sqrt{n})$. It means that to suppress approximation error to ϵ, it requires a large sample volume in the order of $O(1/\epsilon^2)$, where ϵ is an error value.

2.1.2. Exact Asymptotics

Exact asymptotics [12] approximates relative errors of tail probability against the number of samples n. The relative error of the tail probability of Bernoulli samples is approximated by exact asymptotics as follows:

$$\lim_{n\to\infty} \frac{\mathbb{P}[\hat{\mu}_n \leq x]}{\sqrt{\frac{1-x}{2\pi x n}}\frac{\mu}{\mu-x}e^{-nd(x,\mu)}} = 1. \quad (6)$$

Here, $d(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is KL divergence between two Bernoulli distributions with mean p and q. Range of x is $0 < x < \mu$.

Both the confidence interval and the number of samples to achieve error probability δ do not have closed forms in exact asymptotics. They can be derived with numerical methods such as Newton–Raphson method. Exact asymptotics requires true mean μ to be known.

Exact asymptotics is empirically known to provide a good approximation of tail probability even under small n, although it is asymptotic where $n \to \infty$.

2.2. Upper-bound-based approach

Upper-bound-based approaches evaluate tail probability based on inequality. Upper-bound-based methods can evaluate the worst case of tail probability at arbitrary n.

2.2.1. Hoeffding's inequality

Hoeffding's inequality [13] provides the upper bound of tail probability without knowledge of the true mean or distribution of samples.

$$\mathbb{P}[\hat{\mu}_n \leq \mu - \Delta] \leq e^{-2n\Delta^2}$$
$$\mathbb{P}[\hat{\mu}_n \geq \mu + \Delta] \leq e^{-2n\Delta^2} \quad (7)$$

Table 1: *Summary of tail probability evaluation methods.*

Methods	Approach	Assumption	Characteristics
Central Limit Theorem	Asymptotic	σ is known	Low accuracy
Central Limit Theorem (Student's t)	Asymptotic		Low accuracy
Exact Asymptotics	Asymptotic	Bernoulli μ is known	High accuracy
Hoeffding's inequality	Upper bound		Loose upper bound
Chernoff -Hoeffding's inequality	Upper bound	Bernoulli μ is known	Tight upper bound

An upper bound of the confidence interval with error probability δ can be derived with Hoeffding's inequality as follows:

$$\Delta \leq \sqrt{\frac{\log\frac{2}{\delta}}{2n}} \quad (8)$$

An upper bound of the number of samples required to achieve error probability δ can be obtained as follows:

$$n \leq \frac{1}{2\Delta^2}\log\frac{2}{\delta} \quad (9)$$

Hoeffding's inequality has the least assumption so it can be applied to various cases. On the other hand, the upper bound given by Hoeffding's inequality is loose, so it provides large worst values.

2.2.2. Chernoff-Hoeffding's inequality

Chernoff-Hoeffding's inequality provides a tighter upper bound of tail probability by assuming the Bernoulli distribution.

$$\mathbb{P}[\hat{\mu}_n \leq x] \leq e^{-nd(x,\mu)} \quad (0 \leq x \leq \mu)$$
$$\mathbb{P}[\hat{\mu}_n \geq x] \leq e^{-nd(x,\mu)} \quad (\mu \leq x \leq 1) \quad (10)$$

Chernoff-Hoeffding's inequality does not have a closed-form confidence interval. Thus, a confidence interval is derived with numerical methods such as Newton–Raphson method.

The number of samples to achieve error probability δ is bounded by Chernoff-Hoeffding's inequality as follows:

$$n \leq \frac{1}{d(\mu \pm \Delta, \mu)}\log\frac{2}{\delta} \quad (11)$$

Chernoff-Hoeffding's inequality requires true mean μ to be known.

2.3. Summary

Table 1 summarizes evaluation methods of tail probability. Exact asymptotics is expected to approximate values close to actual tail probabilities. Hoeffding's inequality is expected to be reliable in real data analysis because it does not suffer from approximation errors and it does not assume distribution or knowledge of true parameters.

3. Experimental evaluations

We apply the evaluation methods of tail probability to (1) toy data where the true mean is known, and (2) actual listening test data used for training of MOS predictors[1]. In the experiment using toy data, we aim to grasp the concrete behavior of each

[1]Our source code is availabe at https://github.com/todalab/mos-analysis-interspeech2023.

evaluation method of tail probability. In the experiment using actual listening test data, we aim to analyze the reliability of an existing listening test result in detail based on all evaluation methods of tail probability.

3.1. Evaluation of toy data

To compare the evaluation methods of tail probability, we evaluated (1) relative errors of approximation or upper bound against true tail probability, and (2) the relationship between confidence intervals and the number of samples. In the experiment about relative errors, we evaluated tail probabilities of which sample mean of Bernoulli samples deviated less than 0.7 where the true mean was 0.8. The relative errors were computed by taking a ratio of approximation or upper bound against true tail probability, where the true tail probability could be derived with cumulative binomial distribution: $\mathbb{P}[\hat{\mu}_n \leq \frac{k}{n}] = \Psi(k, n, \mu)$. In the experiment about the relationship between confidence intervals and the number of samples, we evaluated 95% confidence interval by using $\delta = 0.05$ error probability. We used $\mu = 0.8$ true mean, and we derived true standard deviation based on a standard deviation of Bernoulli distribution: $\sigma = \sqrt{\mu(1-\mu)}$ if necessary. We derived the true confidence interval by using an inverse of cumulative binomial distribution: $\Delta = \mu - \frac{k}{n} = \mu - \Psi^{-1}(\delta/2, n, \mu)/n$.

3.1.1. Results of toy data

Figure 1 shows relative errors of the tail probability evaluation methods. The true tail probabilities are shown by a 1.0 relative error. The central limit theorem underestimated tail probabilities. This result followed Berry–Esseen theorem, in that there was a large approximation error when n was small. The relative error of the central limit theorem, however, was not mitigated even if n was increased. Using Student's t-distribution instead of normal distribution slightly mitigated the underestimation. Underestimation of tail probability is not preferable for statistical tests because it would result in misidentifying statistically insignificant as significant due to underestimation of the confidence interval. The exact asymptotics approximated tail probabilities with high accuracy except for extremely small n. In addition, it did not underestimate tail probabilities. Hoeffding's inequality and Chernoff-Hoeffding's inequality showed high relative errors because they were upper-bound-based methods. Chernoff-Hoeffding's inequality had a tighter upper bound of relative errors than Hoeffding's inequality as expected. The relative errors of Chernoff-Hoeffding's inequality had higher values than true tail probability by about $\sqrt{n}$. This could be confirmed with relative differences between formulae of exact asymptotics (Eq. (6)) and Chernoff-Hoeffding's inequality (Eq. (10)).

Figure 2 shows the number of samples to achieve a 95% confidence interval based on tail probability evaluation methods. Note that the score range is $[0, 1]$. The five-grade MOS scale is also placed at the top x-axis for speech researchers. Ground truth confidence intervals were discrete values, and they were not shown as a continuous line. The central limit theorem showed smaller confidence intervals than ground truth for many n values. The exact asymptotics showed a slightly higher confidence interval than the ground truth for all n. Hoeffding's inequality and Chernoff-Hoeffding's inequality showed higher confidence intervals than ground truth. This was expected because they were upper-bound-based methods. The underestimation of confidence intervals by the central limit theorem was not preferable for the statistical test. It is therefore advisable

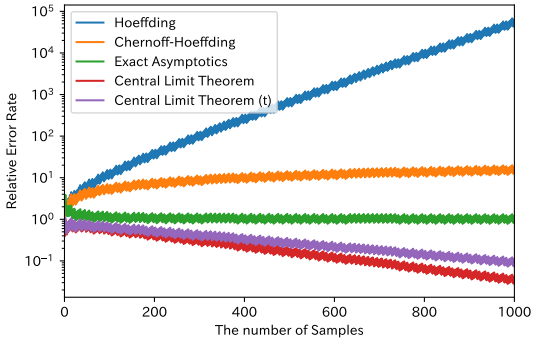

Figure 1: *Relative error of tail probability evaluation methods.*

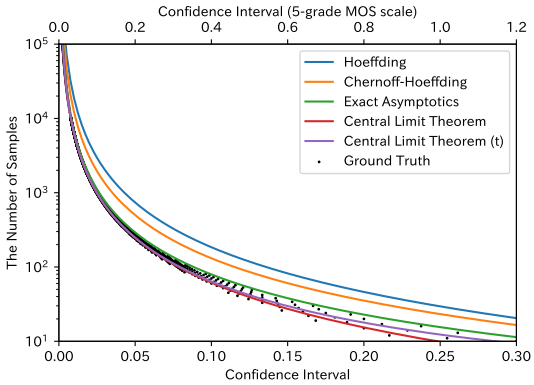

Figure 2: *The number of samples to achieve 95% confidence interval based on tail probability evaluation methods. The x-axis at the top of the figure shows a 5-grade MOS scale.*

to consider that actual confidence intervals could be larger than confidence intervals derived with the central limit theorem.

Table 2 shows concrete values of the number of samples to achieve 95% confidence intervals shown in Fig. 2. As we showed in Fig. 2, common confidence intervals were provided by the central limit theorem with a risk of underestimation, and accurate approximation of confidence intervals was provided by the exact asymptotics, and safe confidence intervals with the least assumption were provided by Hoeffding's inequality. To determine a rank between systems, there must be a difference of twice the confidence interval in the sample mean to be significant, because the lower limit of the confidence interval with the upper system and the upper limit of the confidence interval with the lower system must not be covered. In Table 2, we can see that achieving confidence intervals less than 0.03 requires prohibitive sample volumes considering the cost of crowdsourcing. Recent advancements in the quality of synthesized speech might have reduced the difference in MOS to the level of the limit of the confidence interval that can be ensured with crowdsourcing.

3.2. Listening test data

In the experiment about listening test data, we evaluated confidence intervals of MOS from existing listening test data based on the tail probability evaluation methods. We used a training set of the main task in VoiceMOS challenge [5] as listening test data. This set consisted of five-grade MOS from 175 synthetic speech systems. Each system contains 227 evaluations on average, with a minimum of 96 and a maximum of 288 evaluations per system. If necessary, we substituted true parameters with sample mean and sample standard deviation because true parameters were not known in listening tests.

Table 2: *The number of samples to achieve 95% confidence interval based on tail probability evaluation methods. The "(5-grade MOS)" row shows the confidence interval in the 5-grade MOS scale.*

Confidence Interval	0.0025	0.0075	0.0125	0.025	0.075
(5-grade MOS)	0.01	0.03	0.05	0.1	0.3
Ground truth	100,081	11,094	3,920	1,000	120
Central limit theorem	98,341	10,927	3,934	983	109
Central limit theorem (Student's t)	98,344	10,899	3,936	986	112
Exact Asymptotics	106,141	11,923	4,338	1,113	136
Chernoff-Hoeffding's Inequality	189,459	21,180	7,671	1,946	228
Hoeffding's Inequality	295,110	32,790	11,804	2,951	328

3.2.1. Results of listening test data

Figure 3 shows 95% confidence intervals of MOS based on the tail probability evaluation methods. The systems are sorted from high MOS to low MOS along the x-axis. The central limit theorem showed the smallest confidence interval, and the following methods showed larger confidence intervals in order: central limit theorem with Student's t-distribution, exact asymptotics, Chernoff-Hoeffding's inequality, Hoeffding's inequality. The central limit theorem and central limit theorem with Student's t-distribution had almost the same confidence intervals, so they appeared to overlap in the figure. The confidence intervals of Hoeffding's inequality showed dependency only on the number of samples because it did not consider mean and variance. The confidence intervals from the other evaluation methods reflected mean and variances: systems with high or low MOS showed small confidence intervals because of low variances, and systems with middle MOS showed large confidence intervals because of high variances. The confidence intervals from every method had a large overlap with confidence intervals from adjacent systems with similar MOS, which indicated the quality of adjacent systems could not be ranked with sufficient probability.

Figure 4 shows the number of systems without significant differences between MOS, or the number of systems included inside the confidence interval from each system shown in Fig. 3. The systems with middle MOS had a large number of systems without significant differences. This was because the MOS of many systems was concentrated in the middle range, and systems in this range had relatively large confidence intervals. It also showed that systems with very high MOS had relatively more systems without significant differences than systems with very low MOS. The main reason why many systems in the VoiceMOS challenge had such large confidence intervals and could not determine the rank of adjacent systems was because of the small sample size per system.

Recently, a model-based MOS prediction approach uses the results of a listening test such as the VoiceMOS challenge as training data. Among such methods, there are approaches that predict MOS directly as a training target [7, 8, 10, 14]. As we showed in our experiments, there remained a possibility that indicates sample means or MOS were still distant from the true mean due to the small sample volume. This suggested that MOS evaluated with limited samples might not be appropriate for training data of the automatic quality assessment models. The similar concern was raised by metadata analysis [15].

4. Conclusions and future perspective

This paper evaluated the listening test of synthetic speech with tail probability evaluation methods. We showed that true confidence intervals were expected to be larger than confidence

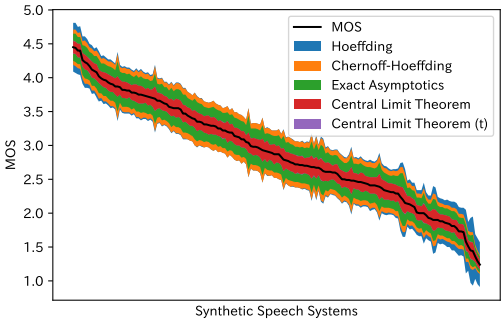

Figure 3: *95% confidence intervals of 175 systems from VoiceMOS challenge.*

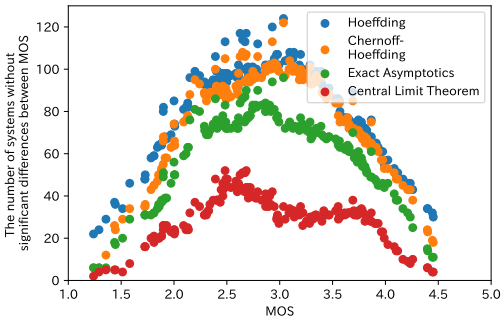

Figure 4: *The number of systems included inside confidence interval from each system from VoiceMOS challenge.*

intervals based on the central limit theorem due to underestimation. We showed that achieving small confidence intervals required large sample volumes that were not collectible with crowdsourcing. We also showed that MOS from existing listening test corpus had large confidence intervals that prevented systems with similar MOS from being ranked with sufficient probability.

A widely-used device for pulling out significant differences from MOS is the biased system selection evaluated in a listening test, but it is ad hoc and makes MOS comparisons across listening tests meaningless. As the fundamental future direction of subjective evaluation, we suggest (1) preferring comparative scoring to direct scoring such as MOS, and (2) optimizing sample assignment of listening tests with online learning. As for (1), it is known that comparative scoring is less noisy than direct scoring [16]. In addition, comparative scoring is recognized to be faster for humans to evaluate [17]. These characteristics of comparative scoring enable a collection of more reliable scores with larger samples under a limited budget. In addition, comparative scoring can avoid calibration issues of MOS [18, 19], which enables the creation of unbiased and combinable training corpora for the quality assessment models. As for (2), preference-based online learning can dynamically select pairs to be compared and optimize the total sample volume to determine the quality of systems under specified accuracy [20]. Comparative scoring has tended to be avoided due to the difficult selection of systems to be compared from the large combination of pairs. However, the ideal criteria to select direct scoring or comparative scoring may not be the number of system combinations but the relative noise level of the scoring methods [16]. Preference-based online learning may enable easy use of comparative scoring with guaranteed accuracy in a listening test.

5. Acknowledgements

This work was partly supported by JST CREST Grant Number JPMJCR19A3.

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, vol. 139. PMLR, 2021, pp. 5530–5540.

[4] S. Shirali-Shahreza and G. Penn, "MOS naturalness and the quest for human-like speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 346–352.

[5] 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.

[6] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 2388–2392.

[7] C. Lo, S. Fu, W. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," in *INTERSPEECH*. ISCA, 2019, pp. 1541–1545.

[8] Y. Leng, X. Tan, S. Zhao, F. K. Soong, X. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *ICASSP*. IEEE, 2021, pp. 391–395.

[9] W. Tseng, C. Huang, W. Kao, Y. Y. Lin, and H. Lee, "Utilizing self-supervised representations for MOS prediction," in *Interspeech*. ISCA, 2021, pp. 2781–2785.

[10] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[11] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941. [Online]. Available: http://www.jstor.org/stable/1990053

[12] Y. V. Prokhorov, "Asymptotic behavior of the binomial distribution," *Uspekhi Mat. Nauk*, vol. 8, no. 3(55), pp. 135–142, 1953. [Online]. Available: http://mi.mathnet.ru/rm8214

[13] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830

[14] W. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *ICASSP*. IEEE, 2022, pp. 896–900.

[15] M. Chinen, J. Skoglund, C. K. A. Reddy, A. Ragano, and A. Hines, "Using rater and system metadata to explain variance in the voicemos challenge 2022 dataset," in *INTERSPEECH*. ISCA, 2022, pp. 4531–4535.

[16] N. B. Shah, S. Balakrishnan, J. K. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright, "When is it better to compare than to score?" *CoRR*, vol. abs/1406.6618, 2014.

[17] N. Stewart, G. Brown, and N. Chater, "Absolute identification by relative judgment," *Psychol Rev.*, vol. Oct;112(4), pp. 881–911, 2005.

[18] E. Cooper, W. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP*. IEEE, 2022, pp. 8442–8446.

[19] 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.

[20] V. Bengs, R. Busa-Fekete, A. E. Mesaoudi-Paul, and E. Hüllermeier, "Preference-based online learning with dueling bandits: A survey," *J. Mach. Learn. Res.*, vol. 22, pp. 7:1–7:108, 2021.