

ACOUSTICAL LETTER

Taut-MUSHRA: A MUSHRA-based method without hidden reference and anchors for relative sound quality evaluation

Fumiyoshi Matano[1,*], Yuya Tagusari[1], Takanori Horibe[1], Junya Koguchi[1] and Masanori Morise[2]

[1]*Graduate School of Advanced Mathematical Sciences, Meiji University, 4–21–1 Nakano, Nakano-ku, Tokyo, 164–8525 Japan*
[2]*School of Interdisciplinary Mathematical Sciences, Meiji University, 4–21–1 Nakano, Nakano-ku, Tokyo, 164–8525 Japan*

(*Received 28 March 2024, Accepted for publication 14 June 2024, J-STAGE Advance published date: 5 July 2024*)

Abstract: State-of-the-art text-to-speech systems have improved in sound quality and have become increasingly large in terms of the number of subjects to detect differences in MOS evaluation, which uses the five-scale precision. The MUSHRA method can precisely detect differences in sound quality compared with the MOS method because sound qualities are rated on a relative scale of 0 to 100 on 101 scales. However, it has the drawback of requiring hidden reference and anchors; thus, it cannot detect cases exceeding the hidden reference. Our method, named *Taut-MUSHRA*, requires no hidden reference and anchors and instead adds two constraints to the subjects. As a result, compared with the MOS method, our Taut-MUSHRA method could more sensitively detect differences in sound quality.

Keywords: Sound quality evaluation, Subjective evaluation, MOS, MUSHRA, Paired comparison method

1. Introduction

In text-to-speech (TTS) synthesis research involving human users, subjective evaluations are essential for comparison with existing methods. Subjects select the sound stimulus they perceive to be of higher sound quality for the presented sound stimulus or rate according to the sound quality. The mean opinion score (MOS) method is a typical evaluation method in which the subjects are asked to rate the sound quality, often based on its naturalness, of a stimulus on a five-point scale. It has the advantage of not requiring a reference, unlike the comparison mean opinion score (CMOS) and degradation mean opinion score (DMOS) methods.

As the sound quality of baselines improves, there is an increasing demand to detect finer differences. For example, in the cases of Tacotron 2 [1], VITS [2], and NaturalSpeech [3], which achieve a quality similar to that of the ground truth, a huge number of subjects is required to discern the differences among methods. It has been pointed out that when the MOS method is compared with the above mentioned methods with such close scores, the sample size required to compare superiority and inferiority is huge [4]. Thus, attempts have been made to predict the MOS by deep learning, but the prediction performance has not yet reached the point where it can estimate a score equivalent to a human [5].

Paired comparison methods such as Thurstone's method offer higher detection performance than the MOS method, but they result in an increase in the number of trials as the conditions being compared expand, potentially imposing a greater burden on subjects [6]. An extended experimental duration may reduce experimental accuracy [7]. Moreover, as only the ranking between methods is known, it is not possible to assess the extent of differences. The MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method [8], which uses multiple scales, is recognized for its high detection sensitivity compared with the MOS method. A Modified-MUSHRA method [9] aims at enhancing detection capability through the introduction of degraded hidden anchors designed to elicit lower scores. However, it may not always be possible to use the ground truth to evaluate TTS systems, as it is not always available for the input text.

In this study, we propose the Taut-MUSHRA, method which combines the high detection power of the MUSHRA method with the simplicity of the MOS method. Instead of the hidden reference and anchors, our Taut-MUSHRA method forces the subjects to give at least one minimum score and one maximum score. This "taut"ens minute differences and is expected to have high detection power among methods with precise differences in sound quality. In evaluation experiments, we compare the MOS method with our Taut-MUSHRA method, results of which indicate the effectiveness of our method.

2. Taut-MUSHRA

Our Taut-MUSHRA method reduces the constraints of the MUSHRA method, which requires the hidden reference and anchors, allowing for the evaluation of sound quality exceeding the baseline. Our Taut-MUSHRA method requires two constraints for the subjects to omit the hidden reference and anchors.

- The score for the sound with the highest squality must be 100, and the score for the sound with the lowest quality must be 0.
- If the sound quality of all stimuli is the same, 100 is assigned to all.

Figure 1 shows the GUI image of our Taut-MUSHRA

*e-mail: matano.fumiyoshi.fx@tut.jp
[doi:10.1250/ast.e24.34]

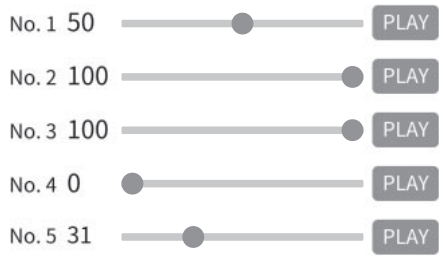

Fig. 1 GUI image of our Taut-MUSHRA method. Subjects are asked to rate the relative sound quality of five speech samples on a scale from 0 to 100 as in the MUSHRA method. In the Taut-MUSHRA method, subjects must evaluate speech according to two constraints.

Table 1 Experimental conditions.

Speaker	One male and one female
# of utterances	50
# of evaluators	30 (ages 19 to 24)
Background noise level	20 dB (soundproof room)
Audio I/O	RME ADI-2 Pro FS R
Headphones	SONY MDR-M1ST

method. All stimuli for comparison are presented, and the subject can listen to them at any time by pressing the "PLAY" button. In each trial, it is examined whether either constraints is met, and if not, a dialog is displayed prompting the subject. This allows experiments to be carried out in the same way as in the MOS method, even under experimental conditions where there is no ground truth.

3. Experiment

In this study, the sound quality evaluation of a downsampled speech with clearly superior or inferior sound quality was conducted to demonstrate the validity of the Taut-MUSHRA method. Table 1 shows the experimental conditions. The speech stimuli used in the experiment consisted of five utterances each from a set of English male speech [10] and Japanese female speech [11], totaling 10 utterances. These were downsampled under five conditions to create a total of 50 speech stimuli. The sound pressure level of a stimulus was determined before the evaluation, and we did not allow the subjects to control the volume during the evaluation. The downsampling in this evaluation affected the sound pressure level negligibly.

We compared our Taut-MUSHRA method with the MOS method. In the MOS method, listening to and evaluating a randomized speech stimulus once counted as one trial, and a total of 50 trials were conducted. In our Taut-MUSHRA method, listening to and ranking five different speech stimuli created from one speech counted as one trial, and a total of 10 trials were conducted. There were no time limits for responses in both methods.

4. Results

Figure 2 shows the results of the MOS and Taut-MUSHRA methods from 29 out of 30 subjects; one was excluded owing to incomplete responses. Error bars represent 95% confidence intervals. From the overall analysis, no significant differences among the results at 96 kHz, 48 kHz, and 32 kHz were observed in the MOS and Taut-MUSHRA methods. Therefore, we discuss the difference between 32 kHz and 24 kHz, which was confirmed in both methods, to examine the ability to detect the differences in sound quality.

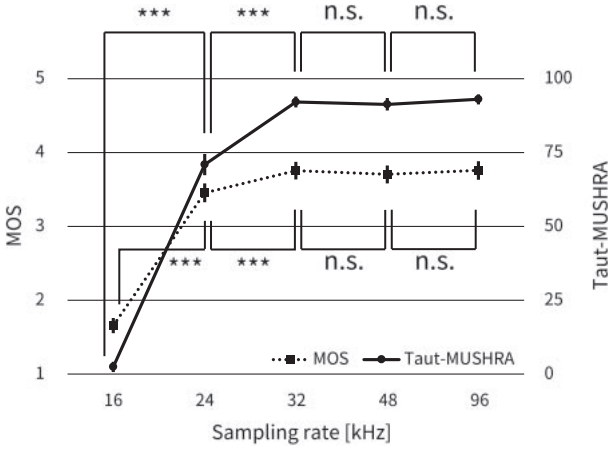

Fig. 2 Results of subjective evaluation. The dashed line represents the results of the MOS method and the solid line represents those of our Taut-MUSHRA method. Error bars represent the 95% confidence interval. The symbol *** represents $p < 0.001$.

First, to verify whether the results at 32 kHz and 24 kHz followed a normal distribution in each method, the Shapiro–Wilk test was conducted, and all results showed $p < 10^{-11}$. Consequently, the Wilcoxon signed-rank test, a nonparametric test, was conducted to examine if there were significant differences between the results at 32 kHz and 24 kHz in each method. For the statistical analysis of the MOS evaluation, the difference between 32 kHz and 24 kHz showed $p = 0.05 \times 10^{-7} \leq 0.05$, and for that of our Taut-MUSHRA method, the difference between 32 kHz and 24 kHz showed $p = 0.04 \times 10^{-34} \leq 0.05$.

We calculated the effect size using the absolute value of Cliff's delta [12], which does not assume a specific distribution. The MOS evaluation showed an effect size of in 0.155, whereas our Taut-MUSHRA method showed 0.356. This indicates that with the same group of subjects, our Taut-MUSHRA method can more significantly detect differences in sound quality. Our Taut-MUSHRA method, as the name suggests, has been shown to be effective in facilitating the detection of small differences by making the MOS results more "taut."

5. Discussion

The distinctive features of our Taut-MUSHRA method are as follows; (1) when a difference in sound quality is detected, scores can range from 0 to 100, with 100 being the highest and 0 the lowest score possible, and (2) when no difference in sound quality is detected, all scores are set to

100. Without these conditions, the concern is that the ability to detect minor differences might decrease, as nearly identical scores would be given for small differences.

Setting the score to 100 when no difference in sound quality is detected suggests that the distribution of evaluation results may not follow a normal distribution. Particularly in evaluations where differences in sound quality are mixed (present and absent), scores of 0 and 100 would coexist. Thus, the use of nonparametric tests becomes a prerequisite when employing this method. Despite these constraints, the findings indicate the possibility of obtaining experimental results with a level of reliability comparable to the MOS method but with a smaller number of subjects.

6. Conclusion

In this study, we examined a subjective evaluation method that ranks speech stimuli, without requiring of hidden reference and anchors in the MUSHRA method. In the experiment, downsampled speech stimuli were used for evaluation, and result of the MOS and Taut-MUSHRA methods were compared. The results showed that differences could be detected more sensitively by the Taut-MUSHRA method than by the MOS method in the same subjects. In the future, we will demonstrate the effectiveness of subjective evaluation using speech generated by modern TTS systems, which can synthesize speech with quality close to the ground truth.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in Aid for Scientific Research grant numbers JP21H04900 and JP22KJ2855.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. S. Ryan, R. A. Saurous, Y. Agiomyrgiannakis and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) '18*, pp. 4779–4783 (2018).

[2] J. Kim, J. Kong and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *Proc. 38th Int. Conf. Machine Learning*, pp. 5530–5540 (2021).

[3] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao and T.-Y. Liu, "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–12 (2024).

[4] Y. Yasuda and T. Toda, "Analysis of mean opinion scores in subjective evaluation of synthetic speech based on tail probabilities," *Proc. Interspeech 2023*, pp. 5491–5495 (2023).

[5] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for Voice conversion," *Proc. Interspeech 2019*, pp. 1541–1545 (2019).

[6] The Japanese Psychonomic Society, *Kiso Shinri-gaku Jikken Handbook* (Asakura Publishing, Tokyo, 2018), p. 160 (in Japanese).

[7] M. Morise, *Speech Analysis and Synthesis* (Corona Publishing, Tokyo, 2018), p. 239 (in Japanese).

[8] International Telecommunication Union, "BS.1534: Method for the subjective assessment of intermediate quality level of audio systems," https://www.itu.int/rec/R-REC-BS.1534/en (accessed 7 Jun. 2024).

[9] M. Rahme, P. Folkeard, S. Beaulac, S. Scollie and V. Parsa, "Modified multiple stimulus with hidden reference and anchors-gabrielsson total impression sound quality rating comparisons for speech in quiet, noise, and reverberation," *J. Speech Lang. Hear. Res.*, **66**, 3677–3688 (2023).

[10] J. Yamagishi, C. Veaux and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," https://datashare.ed.ac.uk/handle/10283/3443 (accessed 20 Mar. 2024).

[11] No. 7 production committee, "No. 7 speech/singing databases," https://voiceseven.com/ (accessed 20 Mar. 2024).

[12] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychol. Bull.*, **114**, 494–509 (1993).