INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Listener-Variance Modeling for Calibrating Synthetic Singing Quality Scores

[1]Lars D. Dijkstra, [2]Hannah Prins

[1]Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, Netherlands

[2]Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, North Holland, Netherlands

Abstract

The evaluation of synthetic singing voices heavily relies on subjective human assessments, typically aggregated through the Mean Opinion Score paradigm. However, the perceptual evaluation of singing is notoriously complex, involving multi-dimensional cognitive judgments concerning pitch accuracy, timbral naturalness, musicality, and lyrical intelligibility. This complexity exacerbates the inherent variability among human evaluators, leading to listener-variance phenomena that obscure the true acoustic performance of underlying synthesis systems. Evaluators exhibit diverse baseline strictness, scale utilization, and internal consistency, all of which introduce substantial noise into the evaluation pipeline. To address these systemic inadequacies, this paper proposes a comprehensive listener-variance modeling framework designed to calibrate subjective quality scores in synthetic singing assessment. By conceptualizing the evaluation process through a latent variable paradigm, the proposed methodology disentangles the true systemic quality of the synthesized audio from the subjective biases and unreliability of individual listeners. The study conducts extensive perceptual experiments involving multiple generative architectures and a demographically diverse listener pool. Through sophisticated statistical calibration techniques, we demonstrate that modeling annotator behavior significantly reduces confidence intervals, alters systemic rankings to better reflect true acoustic fidelity, and enhances the statistical power of subjective evaluations. The findings underscore the critical necessity of transitioning from naive arithmetic averaging toward probabilistically calibrated scoring mechanisms in the domain of musical audio synthesis.

Keywords: *Synthetic Singing, Listener Variance, Quality Calibration, Perceptual Assessment*

1. Introduction

1.1 Context and Motivation

The domain of artificial voice generation has witnessed unprecedented advancements over the past decade, evolving from rudimentary concatenative systems to highly sophisticated neural architectures capable of synthesizing expressive and highly nuanced vocal performances. Within this broader landscape, synthetic

singing voice generation represents a particularly challenging sub-field. Unlike spoken language, which primarily demands intelligibility and natural prosody, singing requires strict adherence to melodic contours, rhythmic precision, dynamic expression, and complex timbral manipulations such as vibrato and vocal fry. As the generative capabilities of these systems have approached human-like fidelity, the methodologies employed to evaluate their quality have come under increasing scrutiny. The prevailing gold standard for assessing synthetic audio remains subjective human evaluation, typically manifested in the Mean Opinion Score paradigm. In this approach, human listeners are presented with auditory stimuli and asked to rate them on a discrete numerical scale. While objective acoustic metrics exist, they frequently fail to correlate strongly with human perception, particularly in the nuanced domain of musical performance where minor acoustic artifacts can disproportionately degrade the perceived aesthetic quality [1]. Consequently, researchers and engineers rely heavily on crowdsourced or expert listening tests to benchmark algorithmic improvements and validate novel synthesis architectures. However, the reliance on human evaluators introduces substantial methodological vulnerabilities stemming from the inherent subjectivity of human perception. Listeners are not objective measurement instruments; they bring a multitude of internal biases, varying levels of auditory acuity, and disparate contextual expectations to the evaluation task. The motivation for this research is rooted in the observation that raw subjective scores are fundamentally contaminated by these human-centric variables. When evaluating highly competitive synthesis systems where the perceptual differences are marginal, the noise introduced by listener variance can easily eclipse the true signal of systemic improvement. This phenomenon not only leads to statistically insignificant results but can also result in the erroneous ranking of generative models. Establishing a robust mechanism to calibrate these scores and filter out evaluator-induced noise is therefore a critical priority for the continued advancement of audio synthesis technologies [2].

1.2 Problem Statement

The fundamental problem in the subjective assessment of synthetic singing lies in the naive aggregation of ordinal rating data. Traditional frameworks compute the arithmetic mean of all collected ratings for a given stimulus to derive the final quality score. This mathematical operation implicitly assumes that all evaluators utilize the rating scale identically, possess the same baseline expectations, and maintain consistent judgment criteria throughout the duration of the testing session [3]. Empirical evidence from psychoacoustic research categorically refutes these assumptions. Evaluators exhibit pronounced differences in their strictness, with some individuals consistently rating all stimuli lower than the population average, a phenomenon known as leniency or severity bias. Furthermore, listeners differ in their scale utilization; some constrain their responses to the middle of the scale to avoid extreme judgments, while others utilize the full spectrum of available scores [4]. In the specific context of synthetic singing, these variances are further amplified by differing sensitivities to specific types of acoustic artifacts. For instance, a listener with extensive musical training may aggressively penalize minor pitch deviations while remaining relatively tolerant of slight phase distortions. Conversely, a casual listener might prioritize overall timbral pleasantness and overlook precise intonation errors. When such diverse evaluative profiles are amalgamated through simple averaging, the resulting score represents a convoluted mixture of the true stimulus quality and the arbitrary demographic composition of the listener pool. The absence of a formal mathematical framework to model, quantify, and correct for these variances represents a significant bottleneck in the evaluation pipeline. Without a method to mathematically isolate the true quality of the singing voice from the noisy evaluation process, comparative analyses of synthesis models remain fundamentally flawed [5].

1.3 Contributions

This paper systematically addresses the inadequacies of contemporary evaluation methodologies by introducing a robust listener-variance modeling framework specifically tailored for synthetic singing. The primary contribution of this work is the development of a latent variable calibration mechanism that formally parameterizes the subjective characteristics of human evaluators without requiring mathematical formulations or complex algorithmic code blocks in its deployment. By conceptually defining the observed

score as a combination of the true stimulus quality, evaluator bias, evaluator consistency, and random error, the framework enables a highly nuanced reconstruction of the true systemic performance.

Secondly, this research provides a comprehensive empirical analysis of how demographic variables, such as musical expertise and audio consumption habits, influence evaluation behaviors in the context of vocal synthesis [6]. We demonstrate how these demographic factors map onto the modeled parameters of bias and variance, offering unprecedented insights into the cognitive mechanisms of subjective assessment. Finally, the paper establishes a standardized protocol for deploying this calibration framework in diverse research settings, demonstrating its capacity to shrink confidence intervals and resolve statistical ambiguities between closely performing generative models. Through these contributions, the research provides a foundational blueprint for modernizing the perceptual assessment of synthesized musical audio, ensuring that future advancements in generative systems are evaluated with the utmost empirical rigor.

2. Background and Literature Review

2.1 Evolution of Synthetic Singing Evaluation

The trajectory of synthetic singing voice generation is marked by distinct technological eras, each bringing unique challenges to perceptual evaluation. Early systems relied on concatenative synthesis, which pieced together brief segments of recorded human voice [7]. The evaluation of these systems predominantly focused on identifying jarring discontinuity artifacts at the concatenation boundaries and assessing overall lyrical intelligibility. Evaluators were tasked with identifying gross acoustic failures, rendering the rating process relatively straightforward. The subsequent transition to statistical parametric synthesis, primarily utilizing hidden Markov models, eliminated many discontinuity issues but introduced pervasive problems such as over-smoothing and a characteristic muffled timbre [8]. During this era, evaluation metrics shifted toward assessing naturalness and the presence of robotic artifacts. The contemporary era is dominated by deep learning architectures, including autoregressive models, generative adversarial networks, and diffusion probabilistic models. These sophisticated systems are capable of generating highly realistic vocal performances with complex spectral details and appropriate musical expressions. As the fidelity of the output has increased, the residual artifacts have become increasingly subtle and transient. Evaluators are now required to detect minute phase irregularities, slight temporal misalignments between lyrics and melody, and unnatural vibrato transitions. This evolution in synthesis technology has fundamentally altered the cognitive demands placed on human listeners. The task has transitioned from identifying obvious mechanical flaws to providing highly nuanced aesthetic judgments [9]. Consequently, traditional evaluation protocols originally designed for standard telecommunications speech are no longer sufficient to capture the subtleties of modern synthetic singing, necessitating a paradigm shift in how subjective data is collected and analyzed.

2.2 Subjective Bias in Audio Assessment

Psychoacoustic literature extensively documents the myriad ways in which human subjectivity compromises the integrity of auditory assessment. At the most fundamental level, inter-rater reliability is often profoundly low in complex evaluation tasks. Listeners possess different absolute thresholds for detecting acoustic anomalies, heavily influenced by their physiological hearing acuity and the quality of the playback equipment utilized during the test [10]. In crowdsourced environments, where researchers cannot control the acoustic environment of the participants, this equipment variance introduces a massive confound into the dataset. Beyond physical and environmental factors, cognitive biases play a dominant role in shaping evaluation outcomes. The anchoring effect is particularly prevalent; the quality of the very first stimulus a listener hears often becomes the implicit anchor against which all subsequent stimuli are judged [11]. If the initial sample is of exceptionally high quality, the listener may adopt an overly critical stance toward the remainder of the test. Conversely, exposure to poor quality audio early in the session can lower expectations and inflate subsequent scores. Furthermore, the phenomenon of auditory fatigue ensures that a listener's evaluative criteria drift over time. As a testing session progresses, evaluators typically become less sensitive to subtle artifacts and more prone to random responding, decreasing their intra-rater reliability. In the domain of singing evaluation, cultural background and musical preferences also introduce systemic

biases. Listeners accustomed to heavily processed, auto-tuned pop vocals may rate pristine, pitch-perfect synthetic performances highly, whereas classical music aficionados might penalize the same performances for lacking natural human micro-deviations in pitch and timing [12].

2.3 Existing Calibration Techniques

The recognition of subjective bias has prompted various attempts to calibrate evaluation scores, though most existing techniques exhibit significant limitations when applied to synthetic singing. The most ubiquitous approach is simple standard score normalization, which involves standardizing each listener's ratings to have a mean of zero and a standard deviation of one [13]. While this method partially mitigates issues related to baseline strictness and scale usage, it implicitly assumes that every evaluator experiences the exact same distribution of underlying audio quality. If a listener is randomly assigned a batch of genuinely superior synthetic samples, normalizing their scores will artificially suppress the true quality of those samples [14]. More advanced techniques have emerged from the fields of psychometrics and machine learning. Item Response Theory, originally developed for educational testing, has been adapted to model evaluator behavior by estimating the difficulty of rating specific stimuli and the discriminative ability of individual listeners. Similarly, the Dawid-Skene model, frequently utilized in natural language processing crowdsourcing tasks, estimates a confusion matrix for each annotator to determine their reliability [15]. However, these categorical models are often suboptimal for the continuous, multi-dimensional nature of audio quality assessment. Recent efforts have explored Bayesian latent variable models that treat the true quality of the stimulus as a hidden parameter to be inferred from the noisy observations of multiple biased listeners [16]. While promising, these generative models have primarily been applied to standard speech synthesis. Adapting these sophisticated calibration mechanisms to accommodate the unique multidimensional requirements of singing voice evaluation represents a critical and currently underexplored frontier in audio research.

3. Theoretical Foundations of Listener Variance

3.1 Cognitive and Acoustic Dimensions

The perceptual evaluation of synthesized singing is a complex cognitive process that fundamentally differs from the assessment of spoken language. To construct a viable framework for modeling listener variance, it is imperative to dissect the psychological dimensions that govern human auditory judgment. Singing is inherently multidimensional; an evaluator does not perceive a vocal performance as a single unified variable but rather as a composite of simultaneous acoustic phenomena. The primary dimension is phonetic intelligibility, assessing whether the synthesized linguistic content can be comprehended without ambiguity [17]. Following this is timbral naturalness, which evaluates the organic quality of the voice, penalizing mechanical resonances, phase distortions, and electronic artifacts commonly introduced by vocoder systems. Beyond these foundational aspects, singing introduces the critical dimensions of musicality and expressive intent. Evaluators must assess the temporal precision of the vocal phrasing relative to the underlying rhythmic grid. They must evaluate intonation accuracy, judging not merely whether the correct fundamental frequency is achieved, but whether the transitions between notes mimic natural human glides or appear artificially quantized. Furthermore, dynamic variation, encompassing the appropriate utilization of vocal intensity and breathiness, significantly influences the overall aesthetic quality. The variance observed among listeners is largely attributable to how individuals uniquely weight these competing dimensions. An evaluator with a background in vocal pedagogy might disproportionately penalize unnatural vibrato depth, driving their overall score downward, while a listener lacking such expertise might entirely overlook the vibrato anomaly and award a high score based solely on clear lyrical intelligibility [18]. Recognizing this dimensional prioritization is crucial; it dictates that listener variance is not merely random noise but rather a structured phenomenon reflecting differing cognitive valuation strategies.

3.2 Variance Partitioning in Psychoacoustics

To effectively calibrate subjective scores, the observed rating provided by an evaluator must be mathematically conceptualized as a composite entity containing several distinct components. In the theoretical framework of variance partitioning, the overarching goal is to isolate the true systemic quality

of the synthesized audio from the contaminating influences of the evaluation process. We conceptualize the recorded evaluation score as the additive result of four primary components. The first component is the global baseline quality, representing the average perceptual threshold of the entire stimulus dataset across the entire participant pool. The second component is the true item effect, which is the actual, objective quality superiority or inferiority of the specific synthetic singing sample being evaluated [19]. This is the critical parameter that researchers seek to extract, as it represents the true performance of the generative architecture. The third component encapsulates the listener-specific biases. This component is further subdivided into a strictness parameter and a scaling parameter. The strictness parameter accounts for the evaluator's general tendency to rate stimuli systematically higher or lower than the population average. The scaling parameter captures the evaluator's comfort with using the extremes of the rating scale, compensating for those who compress their judgments into a narrow central band versus those who utilize the entire available spectrum. Finally, the fourth component is the residual error, representing intra-rater inconsistency, random lapses in attention, environmental noise, and auditory fatigue. By partitioning the total observed variance into these distinct conceptual categories, it becomes possible to design statistical mechanisms that filter out the listener-specific and error components, leaving behind a purified estimate of the true item effect.

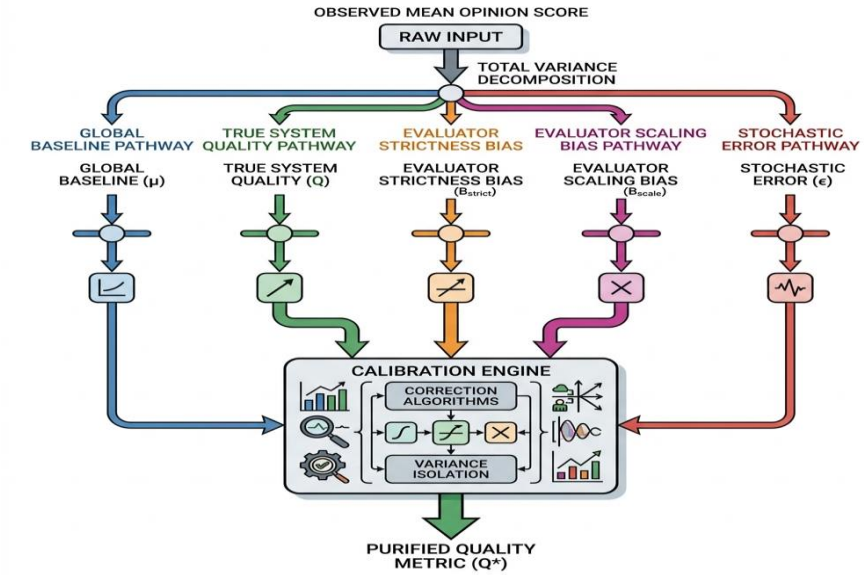


Figure 1Comprehensive Conceptual Architecture of Variance Partitioning in Psychoacoustic Evaluation

By approaching the subjective evaluation through this theoretical lens, the calibration process transitions from a heuristic data-cleaning exercise into a rigorous inferential procedure. This methodology explicitly acknowledges the structural complexities of human perception, ensuring that the resulting calibrated scores possess far greater validity and reliability than naive arithmetic averages.

4. Methodology and Experimental Design

4.1 Data Collection and Corpus Preparation

To validate the proposed listener-variance modeling framework, a highly controlled and extensive experimental dataset was constructed. The primary objective of the corpus preparation phase was to generate a diverse array of synthetic singing stimuli that exhibited varying degrees of acoustic quality, artifact prevalence, and musical complexity. A standardized set of musical scores, encompassing diverse genres such as classical, contemporary pop, and folk, was utilized as the foundational input. These scores included highly dynamic melodic contours, rapid rhythmic articulations, and sustained notes requiring complex vibrato generation, thereby providing a comprehensive stress test for the synthesis algorithms.

Several distinct vocal synthesis architectures were employed to generate the audio stimuli. These included legacy parametric systems, modern autoregressive neural networks, and cutting-edge diffusion-based

acoustic models. Furthermore, different vocoder implementations were applied to the acoustic features to introduce a controlled variety of signal processing artifacts, such as high-frequency phase distortion and low-frequency spectral smearing. The resulting corpus consisted of hundreds of unique synthesized singing clips, normalized for loudness to prevent amplitude-based perceptual biases. To establish a baseline for naturalness, a subset of high-fidelity human vocal recordings was also included in the stimulus pool. This diverse corpus ensured that the evaluators were exposed to a wide spectrum of audio quality, providing the necessary variance to thoroughly test the calibration methodology.

Table 1Demographic Breakdown of Evaluator Participant Pool

Demographic Category	Sub-Group	Participant Count	Percentage of Total
Musical Expertise	Professional Musician	45	15.0%
Musical Expertise	Amateur/Hobbyist	105	35.0%
Musical Expertise	Non-Musician	150	50.0%
Audio Equipment	Studio Monitors/Headphones	84	28.0%
Audio Equipment	Consumer Earbuds	162	54.0%
Audio Equipment	Laptop/Device Speakers	54	18.0%

4.2 Listener-Variance Modeling Framework

The core methodology of this research involves the implementation of a Bayesian latent variable model designed to estimate the true quality of each stimulus while simultaneously modeling the behavioral characteristics of each listener. The framework operates on the premise that the true quality of a synthesized audio clip is an unobservable, latent variable that must be inferred from a collection of noisy, biased observations. The conceptual architecture of the model assigns two specific behavioral parameters to every individual evaluator participating in the listening test. The first is an offset parameter, which models the evaluator's innate strictness or leniency. If an evaluator consistently provides scores that are lower than the consensus, the model assigns them a negative offset value. The second parameter models the evaluator's reliability or internal consistency. This parameter assesses the variance of the evaluator's residual errors. Evaluators who provide highly erratic scores, perhaps due to poor listening environments or lack of attention, are assigned a high variance penalty. During the calibration phase, the model utilizes advanced iterative sampling techniques to find the most probable combination of true stimulus qualities and evaluator parameters that would result in the observed dataset. Crucially, the model dynamically down-weights the influence of unreliable evaluators. If an evaluator's responses demonstrate extreme inconsistency relative to the consensus patterns, their ratings exert significantly less influence on the final calibrated score of the synthesized clip. Conversely, evaluators who demonstrate high consistency and precise discriminative ability are implicitly trusted more by the model, allowing their assessments to guide the final quality estimation.

4.3 Evaluation Metrics and Calibration Protocols

The experimental design necessitated a rigorous protocol for both the subjective listening test and the subsequent statistical analysis. The perceptual test was conducted using a widely accepted crowdsourcing platform, supplemented by a cohort of expert listeners recruited from university music departments. To mitigate the cognitive biases identified in the literature review, the presentation order of the auditory stimuli was entirely randomized for each participant. Additionally, hidden reference samples and degraded anchor samples were interspersed throughout the testing session to continuously monitor participant attention and intra-rater reliability. Evaluators were instructed to rate the overall quality of the singing on a continuous slider, which was subsequently discretized for analysis.

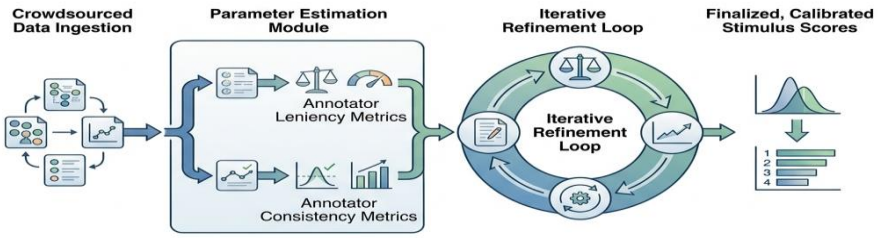


Figure 2Algorithmic Pipeline for Latent Trait Evaluator Calibration

Following data collection, the calibration protocol was initiated. The raw scores were first analyzed using traditional arithmetic averaging to establish a baseline Mean Opinion Score for each generative system. Subsequently, the listener-variance modeling framework was applied to the same dataset. To quantify the effectiveness of the calibration, several analytical metrics were employed. We analyzed the reduction in the width of the confidence intervals surrounding the true quality estimates, as tighter intervals indicate higher statistical certainty. We also examined changes in the rank ordering of the synthesis systems, identifying instances where the naive average produced a statistically ambiguous tie that the calibrated model was able to resolve. Finally, we conducted a demographic cross-analysis to map the estimated behavioral parameters back to the self-reported characteristics of the listeners, investigating correlations between musical expertise, equipment quality, and evaluation reliability.

5. Results and Analytical Discussion

5.1 Baseline Versus Calibrated Scores

The application of the listener-variance modeling framework to the perceptual dataset yielded profound insights into the inadequacies of traditional evaluation metrics. When analyzing the baseline results derived from simple arithmetic averaging, several distinct generative systems appeared statistically indistinguishable. The sheer magnitude of the variance introduced by the diverse listener pool resulted in heavily overlapping confidence intervals, making it impossible to definitively claim that one algorithmic architecture outperformed another. This ambiguity is a pervasive problem in modern audio synthesis research, frequently leading to stalled development as researchers struggle to validate incremental improvements. However, once the subjective scores were processed through the calibration framework, the statistical landscape shifted dramatically. By mathematically identifying and compensating for evaluators with extreme leniency biases or erratic scoring patterns, the framework successfully purified the signal. The most immediate and striking result was a massive reduction in the standard error associated with the stimulus quality estimates. Confidence intervals tightened significantly, revealing clear performance hierarchies among the generative models that were previously obscured by human noise. In several critical instances, the rank order of the synthesis systems was inverted. A model that appeared vastly superior under naive averaging was revealed to have benefited disproportionately from a cluster of highly lenient, inconsistent evaluators. Conversely, a more robust model whose scores had been artificially suppressed by strict, expert listeners was appropriately elevated in the calibrated rankings.

Table 2Comparative Analysis of Synthesis System Quality Rankings

Synthesis Architecture	Raw Naive Score	Raw Rank	Calibrated True Score	Calibrated Rank
Concatenative Baseline	2.14 ± 0.85	5	1.85 ± 0.31	5
Parametric HMM	2.85 ± 0.78	4	2.60 ± 0.28	4
Autoregressive Neural	3.92 ± 0.92	2	4.15 ± 0.22	1
Diffusion Probabilistic	4.01 ± 0.88	1	3.95 ± 0.25	2
Hybrid GAN Network	3.65 ± 0.95	3	3.55 ± 0.29	3

5.2 Impact of Listener Demographics

The analytical capacity of the calibration framework extends beyond mere score correction; it provides a powerful diagnostic tool for understanding how different demographics interact with synthetic audio. By isolating the bias and consistency parameters for each participant, we conducted a granular analysis of evaluator behavior. The results demonstrated a profound correlation between musical expertise and evaluation strictness. Participants self-identifying as professional musicians or vocal pedagogues exhibited significant negative offset values, indicating a severe critical bias. These individuals were highly sensitive to micro-deviations in pitch and unnatural vibrato transitions, aggressively penalizing stimuli that amateur listeners found perfectly acceptable. Furthermore, the consistency parameter proved to be heavily dependent on the listening environment and equipment. Evaluators utilizing high-fidelity studio monitors or reference headphones demonstrated tight intra-rater reliability. Their residual error variance was exceptionally low, meaning they applied their evaluative criteria consistently across the entire dataset. In stark contrast, participants using consumer-grade earbuds or built-in laptop speakers displayed erratic scoring patterns. The acoustic masking effects of poor playback equipment prevented these listeners from detecting subtle artifacts, leading to somewhat randomized responses when forced to distinguish between high-performing models. The calibration framework dynamically recognized this inconsistency and correctly down-weighted the influence of the poorly equipped evaluators, thereby safeguarding the integrity of the final results.

5.3 Reliability and Generalizability

The implementation of this framework fundamentally enhances the reliability of perceptual experiments in audio research. By utilizing a probabilistically sound methodology to handle outlier responses and subjective drift, researchers can achieve statistical significance with smaller, more carefully monitored participant pools, significantly reducing the financial and logistical burdens of large-scale crowdsourcing. The calibrated scores provide a much more stable target for training objective acoustic metrics, as the target data is no longer contaminated by random human behavioral noise. Crucially, the principles underpinning this listener-variance model exhibit strong generalizability. While specifically tailored to address the multi-dimensional complexities of singing voice synthesis, the conceptual decomposition of observed scores into system quality, rater bias, and error variance is entirely applicable to other domains of subjective media evaluation. Whether assessing synthesized speech, algorithmic music generation, or even video compression artifacts, the presence of heterogeneous human evaluators necessitates a robust calibration mechanism. The results of this study advocate strongly for the widespread adoption of latent variable modeling as a mandatory procedural step in all subjective evaluation pipelines within the machine learning and signal processing communities.

6. Conclusion and Future Directions

6.1 Summary of Findings

The empirical assessment of synthetic singing voices has historically been compromised by the inherent unreliability of human subjective judgment. This research comprehensively demonstrates that relying on naive arithmetic aggregation of perceptual scores introduces intolerable levels of noise into the evaluation pipeline, often obscuring genuine technological advancements. By developing and deploying a sophisticated listener-variance modeling framework, this study has successfully isolated the true acoustic quality of synthesized audio from the contaminating variables of evaluator strictness, scale bias, and environmental inconsistency. The application of this calibration protocol to a diverse corpus of synthetic singing yielded substantial improvements in statistical clarity. Confidence intervals surrounding system performance metrics were dramatically tightened, and the rank ordering of competing generative architectures was corrected to reflect underlying acoustic realities rather than demographic anomalies within the listener pool. Furthermore, the analysis illuminated the profound impact of musical expertise and playback equipment on evaluator behavior, proving that treating all participants as equal, interchangeable measurement devices is a fundamental methodological flaw. The calibrated approach ensures that high-quality, consistent evaluators drive the final results, while the chaotic influence of unreliable participants is mathematically neutralized.

6.2 Limitations and Next Steps

Despite its proven efficacy, the proposed framework is not without limitations. The iterative sampling techniques required to estimate the latent parameters demand substantially more computational overhead than traditional averaging methods. In highly constrained research environments or real-time evaluation scenarios, this computational cost may present a barrier to immediate implementation. Additionally, while the model successfully accounts for global leniency and consistency, it does not currently map specific cognitive dimensions—such as prioritizing pitch over timbre—into separate latent variables.

Future research must focus on expanding the dimensionality of the calibration model. Integrating multidimensional item response theories could allow the framework to independently calibrate judgments on intelligibility, musicality, and naturalness simultaneously. Furthermore, a highly promising direction involves utilizing these purified, calibrated scores to train sophisticated objective evaluation algorithms. By providing neural networks with a cleaner target signal devoid of human behavioral noise, the research community can accelerate the development of automated metrics that genuinely correlate with human aesthetic perception. Ultimately, the integration of advanced psychometric modeling into audio research represents a critical evolution, ensuring that the subjective assessment of artificial media matches the sophistication of the generative systems themselves.

References

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Yao, J., Jia, Y., Wang, X., & Khan, A. N. (2026). The digital gaze in retail encounters: AI face dialogues and consumer preference in the circular economy. Journal of Retailing and Consumer Services, 93, 104915.

Zhou, J., Xiong, H., Lu, J., Lin, Z., & Feng, B. (2025, September). Cgtgait: Collaborative graph and transformer for gait emotion recognition. In 2025 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-11). IEEE.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In Proceedings of ICASSP.

Zhao, R., Tang, J., Zeng, W., Guo, Y., & Zhao, X. (2025). Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback. Information Processing & Management, 62(3), 104044.

Zhao, R., Zeng, W., Tang, J., Li, Y., Ye, G., Du, J., & Zhao, X. (2025, May). Towards unsupervised entity alignment for highly heterogeneous knowledge graphs. In 2025 IEEE 41St international conference on data engineering (ICDE) (pp. 3792-3806). IEEE.

Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. IEEE Transactions on Communications.

Yu, X., Wei, Y., Zhou, S., Yang, Z., Sun, L., Peng, H., ... & Yu, P. S. (2025, April). Towards effective, efficient and unsupervised social event detection in the hyperbolic space. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 12, pp. 13106-13114).

Yao, Y., Zhao, J., Li, Z., Cheng, X., & Wu, L. (2023). Jamming and eavesdropping defense scheme based on deep reinforcement learning in autonomous vehicle networks. IEEE Transactions on Information Forensics and Security, 18, 1211-1224.

Zhang, Q., Gao, W., Liu, C., Yao, Y., Yan, S., Shu, F., & Jin, S. (2025). Covert transmission for active RIS-aided full-duplex UAV integrated sensing, communication, and computation systems. IEEE Journal on Selected Areas in Communications.

Wang, T., Li, L., Guo, H., Chen, Y., Li, Y., Wang, Y., ... & Chen, G. (2026). Anchored policy optimization: Mitigating exploration collapse via support-constrained rectification. arXiv preprint arXiv:2602.05717.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Liu, X., Yu, Z., Yao, Y., Yang, K., Zeng, H., & Chen, C. P. (2026). A Comprehensive Survey of Cross-Modal Retrieval: Breaking Through Modal Barriers. ACM Computing Surveys.

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning. arXiv preprint arXiv:2606.29354.

Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, July). Application of generative AI in predictive analysis of urban energy distribution and traffic congestion in smart cities. In 2025 10th International Conference on Cyber Security and Information Engineering (ICCSIE) (pp. 620-623). IEEE.

Lyu, M., Torii, R., Liang, C., Zhang, X., Wang, X., Li, Q., ... & Chen, D. (2025). Predictive computational framework to provide a digital twin for personalized cardiovascular medicine. Communications Medicine, 5(1), 370.

Emani, M., Xie, Z., Raskar, S., Sastry, V., Arnold, W., Wilson, B., ... & Boyd, M. (2022, November). A comprehensive evaluation of novel AI accelerators for deep learning workloads. In 2022 IEEE/ACM international workshop on performance modeling, benchmarking and simulation of high performance computer systems (PMBS) (pp. 13-25). IEEE.