ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464203

A BAYESIAN APPROACH TO SINGING SKILL EVALUATION USING SEMITONE PITCH HISTOGRAM AND MCMC-BASED GENERATED QUANTITIES

Tomoyasu Nakano *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

Conventional approaches to automatic singing skill evaluation have relied either on handcrafted features and their descriptive statistics or on machine learning models trained on large-scale datasets to predict scalar scores such as ratings or rankings. In contrast, we propose a Bayesian modeling approach that estimates probabilistic indicators from a single singing voice, reducing the reliance on large-scale training data when introducing new tasks. As a first step toward a broader framework for singing skill evaluation, we focus specifically on pitch accuracy rather than modeling singing skill as a single comprehensive score for which a commonly used acoustic feature is the *pitch histogram*. This histogram is widely used in reference-independent singing skill evaluation as a descriptor of pitch accuracy. As a variant of this representation, this paper focuses on the *semitone pitch histogram* designed to mitigate variations across musical pieces (scores) by defining pitch bins as subdivisions within plus or minus 0.5 semitones. We introduce a novel Bayesian modeling approach for the semitone pitch histogram. We then estimate its parameters using Hamiltonian Monte Carlo (HMC), enabling probabilistic evaluation of pitch accuracy via generated quantities.

***Index Terms*—** Singing skill evaluation, No-U-Turn Sampler (NUTS), Hamiltonian Monte Carlo (HMC), semitone pitch histogram, generated quantities

1. INTRODUCTION

Singing skill is one of the key elements that characterize the singing voice, which itself is a major component of music. Music plays a central role in both industry and culture, and the global music market has been growing. Many people listen to music with a focus on vocals and lyrics [1], making singing information processing [2] technologies highly beneficial to a wide audience [3]. Technological innovations in this domain, particularly those leveraging deep learning, have thus been actively pursued [4].

Singing-skill-related technologies have seen widespread adoption across both professional and consumer domains. Technologies for correcting pitch are commonly used in popular music production, and scoring functions in karaoke systems have become widely available. On video-sharing platforms such as YouTube, a wide range of individuals, including professionals, publicly share their singing performances. If these performances can be analyzed to estimate singing skill, it opens up new possibilities for music information retrieval and appreciation.

Previous research on automatic singing skill evaluation or singing quality assessment has primarily focused on predicting a single scalar score (*e.g.,* a 7-point rating or ranking) [4]. With these

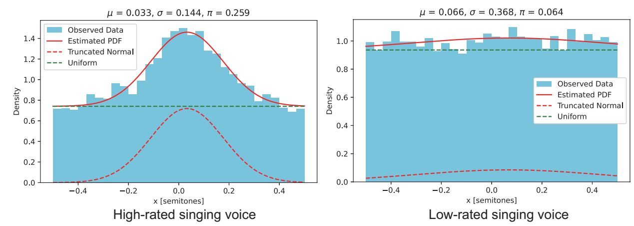

Fig. 1. Semitone pitch histograms of singing voices with high and low singing skill ratings, along with the estimated probability density functions (PDFs) from the proposed mixture model consisting of a truncated normal and a uniform distribution.

approaches, features and their descriptive statistics are either directly used as evaluation indicators, or transformed through linear or nonlinear models trained on large-scale datasets, with or without reference scores. In contrast, this paper aims to develop an approach that outputs *probabilistic indicators* of singing skill from a single vocal performance. The proposed approach is designed to be scalable and adaptable, eliminating the need for large-scale training data when introducing new tasks. It also enables broader applications, including detailed analysis and visualization of individual singing voices and applicability to other sound sources.

We focus on pitch control[1], a crucial aspect of singing skill, and rethink the widely used *pitch histogram* as a model. Specifically, we use Bayesian modeling, with parameter estimation performed via Hamiltonian Monte Carlo, which enables the use of generative quantities for intuitive and probabilistic comparisons, as explained in Section 3. We then investigate how these characteristics relate to singing skill, focusing on pitch accuracy as one component. This approach not only introduces a new perspective to singing skill evaluation but also provides additional insights into singing voice analysis and has the potential to complement and enhance existing methods.

2. SEMITONE PITCH HISTOGRAM

A pitch histogram represents how frequently each pitch is sung throughout a song and has been used in music genre classification [5]. In singing quality assessment, the histogram is typically aggregated over a one-octave range by wrapping the ± 6 semitone bins around the median F0 [6]. Such histograms have been used in machine learning [4, 7, 8] either by extracting features such as the number and positions of peaks, their sharpness, and global histogram statistics (kurtosis and skewness), or directly as acoustic features for deep learning models. With sufficient bin resolution, it can also reflect the pitch accuracy. It is known that higher-skilled singers tend to produce sharper peaks in the histogram, whereas lower-skilled

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number JP25H01174.

[1] In this paper, the fundamental frequency (F0) is treated as pitch.

singing results in a distribution resembling a Gaussian[2] [9].

Such a one-octave pitch histogram enables analysis of pitch-usage patterns and includes information about the frequency of pitch occurrences. However, since the locations and ratios of peaks vary depending on the musical score, this approach inherently reflects song-specific characteristics. In contrast, we emphasize the importance of representations that remain consistent regardless of the particular piece being performed. Therefore, setting the histogram range to one semitone (± 0.5 semitones) yields a more generalizable metric. This approach, referred to as the *semitone pitch histogram*, has also been adopted by Nichols *et al.* [10].

Nichols *et al.* [10] used semitone pitch histograms for singing quality estimation. They employed a histogram divided into seven bins across the semitone range as one of the features for assessing singing quality. They then extracted statistical descriptors from the histogram, such as standard deviation, skewness, and kurtosis. These features were then fed into a passive-aggressive ranker to automatically estimate singing quality based on pairwise comparisons of YouTube videos. This method requires estimating the tuning frequency to align the histogram bins correctly. Gupta *et al.* [9] proposed a set of indicators derived from octave-wrapped pitch histograms, used both as standalone metrics and as machine-learning features, capturing finer shape characteristics and removing the need for tuning-frequency estimation. These indicators were validated against human judgments obtained via best-worst scaling across datasets covering a wide range of singing quality. Furthermore, Nakano *et al.* [11] proposed a related method, known as *semitone stability*, which aggregates pitch information on a semitone scale and is designed to handle differences in tuning frequency.

In comparison to such prior approaches, our contribution lies in proposing a Bayesian framework that leverages generative quantities through Markov chain Monte Carlo (MCMC) sampling. As a first step in this framework, we model semitone pitch histograms to extract interpretable indicators, which are subsequently used both as direct measures and as features for learning-based evaluation. The proposed method is insensitive to moderate tuning frequency and is well-suited for aggregation over short time segments, enabling fine-grained and flexible analysis of singing performance.

3. BAYESIAN MODELING AND MCMC-BASED GENERATIVE QUANTITIES

The No-U-Turn Sampler (NUTS) [12], a variant of Hamiltonian Monte Carlo (HMC), has become easy to use through probabilistic programming languages such as Stan[3]. This also has removed the need for conjugate priors[13], allowing for more flexible and diverse modeling. MCMC-based generative quantities $g(\theta^{(t)})$—defined as functions of samples $\theta^{(t)}$ drawn via MCMC—can serve as the basis for probabilistic inference. For example, the probability that these quantities exceed (or fall below) a reference point c can be computed. In psychology, Toyoda introduced this probability as PHC—the "Probability that the Hypothesis is Correct"—and proposed it as a more intuitive and reproducible alternative to traditional significance testing [13]. Such probabilities, based on the Bayesian posterior, have also been applied in cognitive science [14]. We apply this concept to the interpretation of acoustic features.

3.1. Modeling of Semitone Pitch Histograms

For modeling semitone pitch histograms, the F0 (f_0) in Hertz is converted into semitone units using the following formula:

$$s = 12 \cdot \log_2 \left(\frac{f_0}{f_{\text{ref}}} \right) + 69, \tag{1}$$

where the reference tuning frequency (f_{ref}) is 440 Hz. The semitone value is then wrapped within ± 0.5 semitones using

$$x = (s + 0.5) \bmod 1 - 0.5. \tag{2}$$

This wrapping ensures that correctly sung F0s are centered around zero. To model this behavior, we adopt the truncated normal distribution $\mathcal{N}_T(x; \mu, \sigma^2)$, truncated to the interval $[-0.5, 0.5]$, as it provided stable parameter estimation in our preliminary trials[4]. Yet, three primary factors can cause deviations from zero.

1. Regardless of singing skill, F0 transitions between notes naturally result in non-zero values.
2. Skilled singers often include vibrato, which also leads to non-zero values.
3. Less skilled singers tend to produce non-zero values due to inaccuracies in F0.

Although theoretical models such as sigmoid functions for pitch transitions [15] and sinusoidal functions for vibrato are conceivable in the context of singing voice synthesis and related fields, we do not explicitly model either; instead, we simply assume that all three components are uniformly distributed within their respective domains. Therefore, we adopt a single uniform distribution $\mathcal{U}(x; -0.5, 0.5)$ and then propose the following mixture model.

$$p(x) = \pi \cdot \mathcal{N}_T(x; \mu, \sigma^2) + (1 - \pi) \cdot \mathcal{U}(x; -0.5, 0.5), \tag{3}$$

where $\pi \in [0, 1]$ is the weight of the stable-pitch component.

We place the following priors on the model parameters:

$$\mu \sim \mathcal{N}(0, 0.1^2), \tag{4}$$

$$\sigma \sim \text{Half-Student-}t(3, 0, 0.15), \tag{5}$$

$$\pi \sim \text{Beta}(1, 1), \tag{6}$$

where, μ is the mean of the truncated normal distribution, and we assume it to be flexible enough to accommodate variations in tuning frequency. The standard deviation σ is assigned a Half-Student-t prior, allowing for heavier tails while ensuring positivity. The mixture weight π is given a non-informative $\text{Beta}(1, 1)$ prior.

Figure 1 shows the semitone pitch histograms of singing voices with high and low singing skill ratings. The histogram differences indicate that π and σ can be informative for evaluating singing skill.

3.2. Parameter estimation

To estimate the parameters of the proposed model, we used the Python package CmdStanPy[5] with NUTS. The number of burn-in samples was set to 3000, the number of draws to 1000, and the number of chains to 4. To assess convergence, we used the convergence diagnostic $\hat{R} < 1.01$ and effective sample size (ESS) > 400, as proposed by Vehtari *et al.* [16].

Empirically, we found that this setting led to convergence in most cases with the datasets used in this paper (see Section 4). While

[2] The tendency toward a normal distribution is considered to result from the use of a ± 6 semitone range around the median F0 of the song when calculating the pitch histogram, as singers with lower singing skill often fail to adequately follow the musical score.

[3] `https://mc-stan.org/`

[4] We also considered non-mixture alternatives: the (truncated) normal, von Mises, generalized normal, skew-normal, and sinh-arcsinh distributions.

[5] `https://mc-stan.org/cmdstanpy/`

the proposed model only partially captures aspects of pitch-related skill, the primary aim of this paper is not to guarantee convergence in every scenario, but rather to demonstrate that the model performs reliably with a modest number of samples (iterations).

3.3. Posterior Samples and Generated Quantities for Evaluation

The parameters estimated using the model in Eq. (3) yield a total of $T = 4000$ posterior samples, obtained from 1000 samples per chain across four chains. These samples facilitate inferential statistical analysis through generated quantities such as expectations and credible intervals. While expected a posteriori (EAP) and maximum a posteriori (MAP) estimates provide straightforward means of comparing parameters across performances, the use of generated quantities enables more probabilistic and nuanced analyses that extend beyond simple point estimates. One such quantity is the probability that observations are concentrated around the mean μ. This is computed using the following formulation at each iteration t:

$$p_\delta^{(t)} = \pi^{(t)} \cdot \frac{\Phi(\delta; \mu^{(t)}, \sigma^{(t)}) - \Phi(-\delta; \mu^{(t)}, \sigma^{(t)})}{\Phi(0.5; \mu^{(t)}, \sigma^{(t)}) - \Phi(-0.5; \mu^{(t)}, \sigma^{(t)})}, \quad (7)$$

where $\Phi(\cdot; \mu^{(t)}, \sigma^{(t)})$ denotes the cumulative distribution function (CDF) of the normal distribution with mean $\mu^{(t)}$ and standard deviation $\sigma^{(t)}$, and the denominator is the normalization constant for the truncated normal component. For example, setting $\delta = 0.1$ yields the probability mass within the interval $[\mu^{(t)} - 0.1, \mu^{(t)} + 0.1]$.

Posterior samples and generated quantities can be used to assess the probability that a hypothesis U holds. Let $U^{(t)}$ denote the logical condition indicating whether U is true for the t-th posterior sample. The indicator function (*i.e.,* generated quantity) $\mathbb{I}(U^{(t)})$ returns 1 if $U^{(t)}$ is true, and 0 otherwise. The resulting probability is given by

$$\mathrm{PHC}(U) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(U^{(t)}). \quad (8)$$

For example, to estimate the probability that p_δ exceeds a threshold c, let $p_\delta^{(t)}$ denote the t-th generated sample and define

$$U^{(t)} \equiv p_\delta^{(t)} > c. \quad (9)$$

Then, $\mathrm{PHC}(U)$ represents the probability that p_δ is greater than c. PHC also applies to other hypotheses, including pairwise comparisons between performances (*e.g.,* $|\pi_A^{(t)} - \pi_B^{(t)}| > c$) and conjunctions of conditions (e.g., $|\theta_B^{(t)} - \theta_A^{(t)}| > c_\theta$ and $|\pi_A^{(t)} - \pi_B^{(t)}| > c_\pi$).

4. EXPERIMENT

We used our in-house dataset introduced in [17], which includes 140 solo singing renditions (20 songs, each sung by 7 singers) of Japanese popular songs. Each rendition was mixed with background music and annotated by ten experts using a 7-point Likert scale across six criteria: pitch, rhythm, pronunciation, expression, vocal projection, and overall performance. We focus only on the evaluation scores for the pitch criterion. The ten pitch scores for each rendition were aggregated into a single value using the EAP estimate derived from the best-performing item response theory (IRT) model proposed in [17]. The model was selected according to the expected log pointwise predictive density [18] as an information criterion.

The singing voices we analyzed were dry solo vocal recordings. All audio files were converted to 16 kHz mono for analysis. F0 was estimated using WORLD [19], and frames identified as unvoiced were excluded from the analysis. Figure 2 shows the distribution of evaluation scores given by ten annotators. The scores are distributed across a wide range, indicating variability in assessments. Most renditions received scores clustered around the middle range, while a few received particularly high or low evaluations.

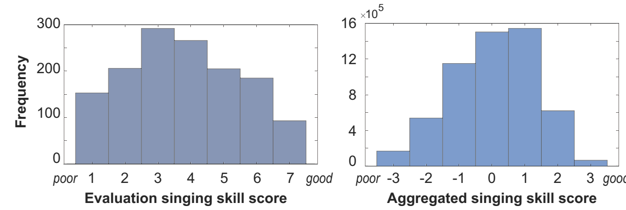

Fig. 2. Left: Histograms of pitch evaluation scores by 10 annotators for 140 singing renditions. Right: Histogram of all posterior samples of aggregated latent scores estimated by the IRT model (140 × 40,000 samples).

Parameter	Converged (%)
σ	123 (87.9%)
π	122 (87.1%)
$p_{\delta=0.05}$	134 (95.7%)
$p_{\delta=0.10}$	132 (94.3%)

Parameter	Converged (%)
$p_{\delta=0.15}$	129 (92.1%)
$p_{\delta=0.20}$	127 (90.7%)
$p_{\delta=0.25}$	125 (89.3%)

Table 1. Number and percentage of converged songs (out of 140) for which both $\hat{R} < 1.01$ and ESS > 400.

4.1. Results

Table 1 shows, for each parameter, the number and percentage of converged singing renditions (out of 140). We argue that in most cases where the convergence criteria are met, the effectiveness of the proposed approach is a meaningful and valid subject of discussion.

Figure 3 shows scatter plots with bars, illustrating the relationship between the aggregated singing skill scores and each of the model parameters (π, σ, $p_{\delta=0.10}$, and $p_{\delta=0.25}$). For each parameter, the MAP and EAP estimates are shown with the 3-97% highest density interval (HDI). The plots suggest some association with the aggregated scores, and the correlation coefficients (EAP / MAP) are 0.34 / 0.43 for π, -0.30 / -0.19 for σ, 0.44 / 0.45 for $p_{\delta=0.10}$, and 0.42 / 0.46 for $p_{\delta=0.25}$.

4.2. Discussion

Based on the results, the proposed π, $p_{\delta=0.10}$, and $p_{\delta=0.25}$ tended to increase as the aggregated score improved. Similarly, σ tended to decrease (*i.e.,* become sharper) as singing skill increased, converging between 0.1 and 0.2 once a certain level of skill was reached. These findings suggest that these parameters can be used to evaluate singing skill from the perspective of pitch. According to the correlation coefficients, p_δ aligns more closely with human evaluations of singing skill than π, which in turn aligns more closely than σ.

To compare the representational abilities of π and p_δ, Figure 4 is presented based on Eq. (8). This figure shows the probabilities (*i.e.,* PHC) that π and $p_{\delta=0.25}$ are equal to or greater than c ranging from 0 to 1.0 in increments of 0.1. At $c = 0$, PHC is equal to 1, whereas at $c = 1$, PHC is equal to 0. Computing PHC enables comparisons that explicitly account for the posterior uncertainty. If PHC remains high even when c is large (shown as yellow near the top of the figure), it means the indicator reflects singing skill well. In

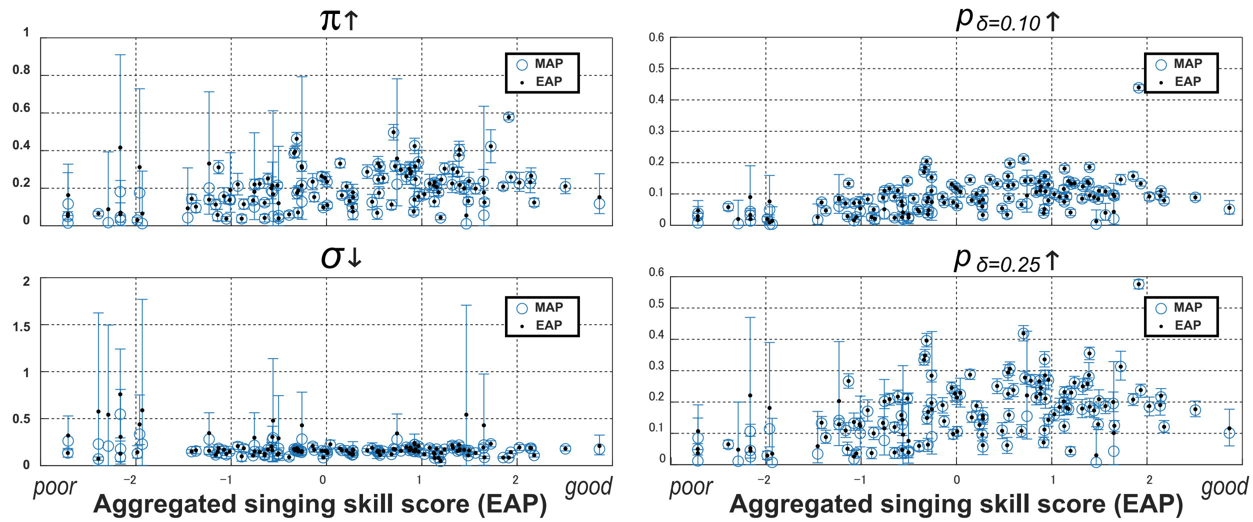

Fig. 3. Scatter plots of aggregated singing-skill scores (EAP) versus model parameters (π, σ, $p_{\delta=0.1}$, $p_{\delta=0.25}$) for converged songs. MAP/EAP parameter estimates are shown with 3-97% HDI bars.

the figure, both π and p_δ show higher probabilities at larger c values for performances on the right side, which have higher singing skill. For low-scoring singing performances, π tends to yield high PHC values even when c is large, whereas p_δ suppresses this tendency, potentially resulting in values that better reflect actual singing skill.

Since π includes F0 transitions between notes that are not directly related to singing skill, removing these components is expected to improve estimation accuracy. Vibrato may also reduce π, thus negatively impacting singing skill estimation. However, simply detecting and removing vibrato would also exclude sustained tones—an important feature for characterizing singing skill—from estimation. Moreover, pitch perception in vibrato singing is known to be ambiguous [20], and indiscriminate smoothing of vibrato introduces further complications. Because vibrato is mentioned in the free-text comments that form the basis of the evaluations, annotators took it into account when assessing pitch-related skills. Therefore, a method to appropriately compensate for vibrato is necessary.

Since the annotations were made under conditions in which the singing was mixed with background music, similar to normal music listening, taking into account the relationships with other instruments can enable more accurate and detailed analysis.

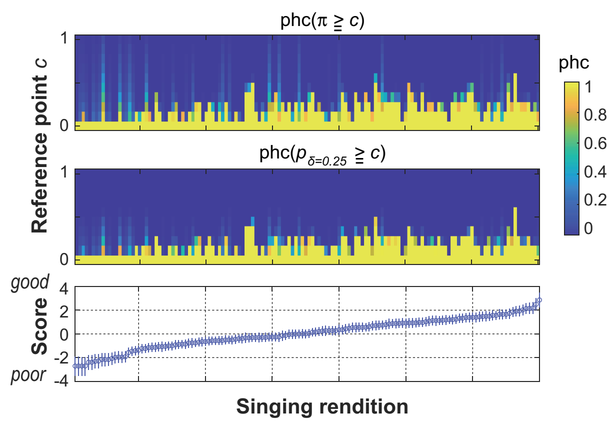

Fig. 4. For songs with converged estimates. PHC that π or $p_{\delta=0.25}$ exceeds c (top, middle; shown as colormaps) and aggregated singing skill score based on pitch criteria (bottom). In the bottom panel, the horizontal axis lists singing renditions sorted by EAP (higher to the right), and error bars show the 3-97% HDI. In high-skilled performances (right side), yellow extending to larger c indicates a high posterior probability that the quantity (*e.g.,* $p_{\delta=0.25}$) is large.

5. CONCLUSION

We define singing skill as an objective ability, specifically the technique of controlling the singing voice. Because human evaluations inherently involve ambiguity, it is important not only to develop automatic evaluation methods based on machine learning from subjective ratings but also to extract consistent, interpretable indicators directly from acoustic features without relying on human judgments.

We modeled the semitone pitch histogram using Japanese popular music as a first step toward understanding singing proficiency. The semitone pitch histogram has a narrow domain and, thanks to Bayesian modeling, can be handled even with small amounts of data. This makes it applicable not only to the analysis of entire songs but also to short-term segments. We believe that this approach can enhance the explainability of singing analysis, for example by identifying which parts of a performance are particularly strong or weak. Looking ahead, we will extend the proposed Bayesian modeling framework by incorporating explicit vibrato modeling and the joint inference of multiple pitch-related features, thereby broadening the analysis to additional aspects of singing performance, including rhythm. Because PHC enables probabilistic comparisons, including (i) the probability that one rendition outperforms another and (ii) the PHC of a conjunctive hypothesis in which multiple indicators for a single rendition exceed preset thresholds, we will also leverage this framework for interaction design and information visualization that support the development of singing skills. Building on these directions, we will further examine the proposed indicators alongside existing approaches and investigate F0 estimation methods suitable for this task, thereby integrating insights on tuning-frequency differences, pitch-related processing (*e.g.,* F0 trajectory analysis), and acoustic features into a more unified framework.

6. REFERENCES

[1] Andrew M. Demetriou, Andreas Jansson, Aparna Kumar, and Rachel M. Bittner, "Vocals in music matter: The relevance of vocals in the minds of listeners," in *Proc. ISMIR 2018*, 2018.

[2] Masataka Goto, Takeshi Saitou, Tomoyasu Nakano, and Hiromasa Fujihara, "Singing information processing based on singing voice modeling," in *Proc. IEEE ICASSP 2010*, 2010, pp. 5506–5509.

[3] Eric J. Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M. Bittner, Andrew M. Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, Anna M. Kruspe, and Luwei Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 82–94, 2019.

[4] Chitralekha Gupta, Haizhou Li, and Masataka Goto, "Deep learning approaches in topics of singing information processing," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 2422–2451, 2022.

[5] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," *Journal of New Music Research*, vol. 32, no. 2, pp. 143–152, 2003.

[6] Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic leaderboard: Evaluation of singing quality without a standard reference," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 13–26, 2020.

[7] Ping-Chen Chan, Po-Wei Chen, and Von-Wun Soo, "Improve singing quality prediction using self-supervised transfer learning and human perception feedback," in *Proc. MMAsia 2023*, 2023, pp. 1–7.

[8] Yaolong Ju, Chun Yat Wu, Betty Cortiñas-Lorenzo, Jing Yang, Jiajun Deng, Fan Fan, and Simon Lui, "End-to-end automatic singing skill evaluation using cross-attention and data augmentation for solo singing and singing with accompaniment," in *Proc. ISMIR 2024*, 2024, pp. 493–500.

[9] Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic evaluation of singing quality without a reference," in *Proc. APSIPA ASC 2018*, 2018, pp. 990–997.

[10] E. Nichols, C. DuHadway, H. Aradhye, and R. Lyon, "Automatically discovering talented musicians with acoustic analysis of YouTube videos," in *Proc. IEEE ICDM 2012*, 2012, pp. 559–565.

[11] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Interspeech 2006*, 2006, pp. 1706–1709.

[12] Matthew D. Hoffman and Andrew Gelman, "The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.

[13] Hideki Toyoda, *Statistics with posterior probability and a PHC curve*, Springer Singapore, 2024.

[14] Kazuhiro Ogata, Reo Gakumi, Atsushi Hashimoto, Yoshitaka Ushiku, and Shigeo Yoshida, "The influence of bouba- and kiki-like shape on perceived taste of chocolate pieces," *Frontiers in Psychology*, vol. 14, pp. 1–13, 2023.

[15] Jordi Bonada and Merlijn Blaauw, "Hybrid neural-parametric f0 model for singing synthesis," in *Proc. IEEE ICASSP 2020*, 2020, pp. 7244–7248.

[16] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner, "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC," *Bayesian Analysis*, pp. 1–38, 2021.

[17] Tomoyasu Nakano and Masataka Goto, "Using item response theory to aggregate music annotation results of multiple annotators," in *Proc. ISMIR 2024*, 2024, pp. 1–9.

[18] Aki Vehtari, Andrew Gelman, and Jonah Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[19] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[20] Johan Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, 1987.