Online Preference Elicitation for Personalized Singing Synthesis: A Comprehensive Framework for User-Centric Acoustic Adaptation

Joseph Smith[1], Abigail Martinez[2]

[1] School of Communication Sciences and Disorders, Dalhousie University, Halifax, Nova Scotia, Canada
[2] School of Communication Sciences and Disorders, Dalhousie University, Halifax, Nova Scotia, Canada

Abstract

The rapid advancement of artificial intelligence in audio generation has significantly improved the naturalness and intelligibility of singing synthesis systems. However, a critical gap remains in aligning the generated vocal performances with the highly subjective and individualized preferences of different users. Conventional systems rely on static datasets and generalized models that output standardized acoustic features, largely ignoring idiosyncratic tastes regarding vibrato depth, breathiness, tension, and emotional delivery. This paper presents a comprehensive conceptual framework and methodology for online preference elicitation designed specifically for personalized singing synthesis. By integrating human-in-the-loop machine learning techniques, the proposed system dynamically interacts with the user, presenting iteratively generated vocal samples and capturing feedback in real time to update an underlying preference model. We explore the parameterization of acoustic features relevant to vocal aesthetics and detail the active learning strategies employed to minimize user fatigue while maximizing the convergence rate of the preference vector. Through detailed architectural design and extensive experimental validation involving both subjective perceptual evaluations and objective convergence metrics, we demonstrate that online preference elicitation significantly outperforms static personalization methods. The findings indicate that users achieve a higher degree of satisfaction and perceived control over the synthesis output. This research bridges the gap between sophisticated generative audio models and human-computer interaction, paving the way for intuitive, accessible, and deeply personalized creative tools in music production.

Keywords: Personalized Singing Synthesis; Online Preference Elicitation; Active Learning; Human-Computer Interaction

1. Introduction

1.1 Motivation and Context

The domain of artificial intelligence in music generation has witnessed a paradigm shift over the past decade, evolving from rudimentary robotic vocalizations to highly expressive and indistinguishable synthetic singing voices. This evolution has been primarily driven by the advent of deep learning architectures capable of modeling the complex temporal and spectral dependencies inherent in human vocal production. Current singing synthesis systems can generate highly realistic vocal tracks given a musical score and corresponding lyrics. These systems are increasingly utilized in various applications, ranging from virtual idols and automated music production to assistive technologies for individuals with vocal impairments. Despite these profound technical achievements, the widespread adoption of singing synthesis as a flexible creative tool is hindered by a fundamental limitation regarding user personalization. Most contemporary models are trained to optimize a generalized objective function over a large corpus of vocal data, resulting in a one size fits all vocal output that reflects the average characteristics of the training dataset rather than the specific artistic vision of an individual creator. The perception of a perfect vocal performance is inherently subjective and varies drastically among producers, composers, and listeners as noted in early perceptual studies [1]. Some users may prefer a highly controlled and bright vocal tone with precise pitch transitions, while others might favor a breathy, relaxed performance with delayed vibrato onset and slight intentional pitch deviations. Conventional systems require users to possess deep domain knowledge in digital signal processing to manually tweak low level acoustic parameters to achieve these desired stylistic nuances. This manual manipulation is highly unintuitive, time consuming, and often disrupts the creative workflow. Therefore, there is a pressing need for intelligent mechanisms that can automatically adapt the synthesis output to the subjective tastes of the user without requiring explicit technical expertise [2].

1.2 The Need for Online Preference Elicitation

To address the challenge of personalization, some researchers have proposed offline adaptation techniques such as fine tuning a pre trained model on a small dataset of target vocals. While effective for speaker voice cloning, offline adaptation is ill suited for capturing stylistic and aesthetic preferences, as these preferences

are often abstract, context dependent, and difficult for users to articulate or provide explicit training data for. Furthermore, a user might not know exactly what vocal style they are looking for until they hear various alternatives. This phenomenon highlights the necessity for an exploratory and interactive approach to personalization. Online preference elicitation emerges as a promising solution to this problem. Originating from the fields of decision theory and recommender systems, online preference elicitation involves dynamically querying a user to gather information about their preferences and continuously updating a model of their utility function in real time [3]. In the context of singing synthesis, this translates to presenting the user with a sequence of synthesized vocal samples, each varying slightly in acoustic characteristics, and asking the user to provide feedback. This feedback can take the form of pairwise comparisons, ratings, or directional adjustments. By employing advanced active learning algorithms, the system can intelligently select the most informative samples to present to the user at each step, thereby rapidly narrowing down the search space of acoustic parameters to find the optimal configuration that maximizes the user satisfaction [4]. This interactive loop not only democratizes the use of complex synthesis engines by replacing technical sliders with intuitive auditory comparisons but also accommodates the fluid and exploratory nature of the musical creation process.

1.3 Research Objectives and Contributions

The primary objective of this research is to design, implement, and evaluate a comprehensive framework for online preference elicitation in the domain of personalized singing synthesis. We aim to demonstrate that integrating human in the loop machine learning techniques with state of the art vocoders and acoustic models can yield highly customized vocal performances that align with abstract user preferences. To achieve this, we first establish a robust parameterization scheme that maps high level perceptual qualities to controllable low level acoustic features [5]. Second, we develop a Bayesian optimization based elicitation strategy that efficiently queries the user while managing the inherent cognitive load associated with auditory evaluation. Third, we architect a complete end to end system that facilitates real time interaction and seamless synthesis updates. Finally, we conduct rigorous subjective and objective evaluations to validate the efficacy of our proposed framework against traditional static interfaces. The contributions of this paper are therefore multi faceted, offering theoretical insights into the modeling of auditory preferences, practical engineering solutions for real time interactive audio synthesis, and empirical evidence supporting the superiority of active elicitation methods in creative applications. The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of the relevant literature. Section 3 outlines the conceptual framework and methodology. Section 4 details the implementation and system architecture. Section 5 presents the experimental setup, results, and discussion. Section 6 concludes the paper and outlines directions for future research.

2. Literature Review and Background

2.1 Evolution of Singing Synthesis Systems

The history of singing synthesis is characterized by a continuous effort to capture the physiological and acoustic complexities of the human voice. Early approaches relied heavily on concatenative synthesis, where short segments of recorded human vocals were stored in a large database and stitched together to form new sentences or melodies [6]. While concatenative systems could preserve the natural timbre of the original singer, they suffered from noticeable artifacts at the concatenation boundaries and lacked the flexibility to modify the expressive dynamics of the performance. The introduction of statistical parametric synthesis marked a significant milestone. These systems utilized hidden Markov models to model the trajectory of acoustic features such as fundamental frequency and spectral envelopes [7]. Statistical parametric synthesis offered greater control over the acoustic parameters and resulted in smoother transitions between phonemes. However, the generated audio often suffered from a characteristic muffled or buzzy quality due to the oversimplification of the phase information and the constraints of the underlying vocoders. The breakthrough in singing synthesis arrived with the application of deep neural networks. Autoregressive models demonstrated the capability to generate raw audio waveforms with unprecedented fidelity by modeling the conditional probability of each audio sample given the previous samples [8]. Despite their high quality, autoregressive models were computationally expensive and unsuitable for real time applications. Subsequent advancements led to the development of non autoregressive architectures, which parallelized the generation process and significantly reduced inference time [9]. These modern acoustic models, coupled with neural vocoders, currently represent the state of the art, capable of producing highly natural and expressive singing voices. However, as these models have grown in complexity, their internal representations have become increasingly opaque, making it difficult for end users to intuitively manipulate the stylized output.

2.2 Personalization in Audio Generation

As the base quality of synthesized singing reached a plateau of high naturalness, research focus shifted towards personalization and expressivity control. Early attempts at personalization primarily dealt with speaker adaptation, aiming to mimic the vocal timbre of a target singer using a minimal amount of training data. Techniques such as speaker embedding vectors and few shot learning have been extensively explored in text to speech and subsequently adapted for singing synthesis. These methods allow a single multi speaker model to generate voices that sound like different individuals by conditioning the acoustic model on a specific latent representation. While highly effective for timbre cloning, these techniques do not address the stylistic preferences of the user. Two users utilizing the exact same cloned voice may desire completely different delivery styles for a given song. To address style control, researchers have proposed various disentanglement techniques, attempting to separate timbre, pitch, rhythm, and emotional content within the latent space of the neural network [10]. By manipulating these disentangled representations, users can theoretically alter the vibrato, breathiness, or tension of the generated voice. However, these systems typically provide users with a complex dashboard of abstract sliders representing the latent dimensions. Mapping a specific musical intention, such as a desire for a more melancholy delivery, to a combination of multidimensional slider adjustments remains a daunting task for the average user. This disconnect between the user intention and the control interface highlights the limitations of purely technical solutions to personalization and underscores the need for user centric interaction paradigms.

2.3 Preference Elicitation and Active Learning

Preference elicitation is a well established subfield of artificial intelligence dedicated to constructing models of user preferences through active querying. It has been widely applied in domains such as personalized medicine, product recommendation, and multi criteria decision making [11]. The core challenge in preference elicitation is to gather the maximum amount of information about the user utility function with the minimum number of queries, as human users are prone to fatigue and inconsistency. Active learning strategies are heavily utilized to achieve this efficiency. Instead of presenting queries randomly, an active learning system analyzes the current uncertainty in the preference model and selects the query that is expected to reduce this uncertainty the most [12]. In recent years, preference elicitation has been integrated with deep learning models in a paradigm known as human in the loop machine learning. This approach has shown remarkable success in areas where the objective function is inherently subjective and difficult to define mathematically, such as generative art and procedural content generation. For instance, interactive evolutionary computation algorithms have been used to guide the generation of visual art by having users select their favorite images from a population of mutated candidates [13]. Applying these concepts to audio generation, however, presents unique challenges. Unlike visual stimuli, which can be evaluated almost instantaneously, audio samples unfold over time. Evaluating a set of singing samples requires the user to listen to each sample sequentially, significantly increasing the cognitive load and the time required for each elicitation step. Therefore, adapting preference elicitation for singing synthesis requires careful consideration of the querying mechanism, the parameterization of the acoustic space, and the design of the user interface to ensure a seamless and efficient experience.

3. Conceptual Framework and Methodology

3.1 The Human in the Loop Synthesis Paradigm

The proposed conceptual framework operates on a cyclical human in the loop paradigm that bridges the gap between the complex parameter space of the singing synthesis engine and the intuitive perceptual space of the user. The primary objective is to iteratively refine a set of acoustic control parameters until the synthesized output aligns with the internal, often unarticulated, aesthetic preference of the user. The process begins with the user providing a standard input consisting of a musical score encompassing notes and durations alongside the corresponding lyrics [14]. An initial set of acoustic parameters is sampled from a prior distribution representing the average singing style. The synthesis engine processes the score, lyrics, and these initial parameters to generate a baseline vocal performance. This audio sample is presented to the user. From this point, the active elicitation loop commences. The system generates one or more alternative performances by intelligently perturbing the acoustic parameters. The user listens to these alternatives and provides feedback based on their subjective preference. This feedback is processed by a preference model, which updates its internal understanding of what the user considers optimal [15]. The updated preference model then informs the selection of the next set of acoustic parameters to be synthesized and evaluated. This iterative loop of synthesis, evaluation, and update continues until the preference model converges to a stable state or the user is satisfied with the result. The success of this paradigm hinges on two critical components which are the interpretable parameterization of acoustic features and the efficiency of the online elicitation strategy.

3.2 Acoustic Feature Parameterization

To facilitate meaningful manipulation and preference modeling, it is essential to define a parameter space that is both acoustically relevant and perceptually interpretable. Direct manipulation of the raw latent vectors of a deep neural network is inefficient for elicitation because these spaces are highly entangled and lack clear acoustic correlates. Therefore, we establish an intermediate control layer that parameters specific aspects of vocal production. These parameters are then mapped to the conditioning inputs of the acoustic model. The parameterization focuses on four primary dimensions of vocal expression which include pitch dynamics, spectral energy distribution, temporal fluctuations, and vocal tension. Pitch dynamics govern the micro tonal variations around the nominal pitch dictated by the score. This includes the depth and frequency of vibrato, the speed of portamento transitions between notes, and intentional pitch drift. Spectral energy distribution relates to the perceived breathiness or clarity of the voice. By adjusting the spectral tilt and the ratio of harmonic to noise components, the system can transition between a whispery, intimate tone and a bright, resonant projection. Temporal fluctuations refer to the micro timing deviations from the rigid musical grid, which impart a sense of groove and human imperfection to the performance [16]. Finally, vocal tension simulates the physiological effort of the singer, manipulating the spectral envelope to sound either relaxed and effortless or strained and powerful. These high level parameters provide a structured and bounded search space for the preference elicitation algorithm.

Table 1: Description of Acoustic Parameters

Parameter Category	Specific Feature	Perceptual Effect on Synthesized Voice
Pitch Dynamics	Vibrato Extent	Controls the amplitude of pitch modulation during sustained vowels
Pitch Dynamics	Portamento Speed	Dictates the transition duration between two distinct consecutive notes
Spectral Energy	Harmonic to Noise Ratio	Modifies the balance between clear voicing and breathy aspiration noise
Spectral Energy	Spectral Tilt	Alters the overall brightness and presence of high frequency overtones
Vocal Tension	Formant Shifting	Simulates physical vocal tract constriction for varying emotional intensity

3.3 Online Elicitation Strategy

The core algorithmic challenge in this framework is determining how to query the user to rapidly accurately deduce their optimal parameter configuration. We formulate this as an interactive optimization problem where the objective function, representing user satisfaction, is unknown and expensive to evaluate since evaluating requires the user to listen to audio. We employ a strategy grounded in Bayesian optimization to navigate this challenge [17]. The system maintains a probabilistic surrogate model, typically a Gaussian process, which maps the acoustic parameter space to an estimated preference score. Initially, the surrogate model has high uncertainty across the entire parameter space. During each iteration, the system uses an acquisition function to select the next set of parameters to synthesize. The acquisition function balances exploitation, which means selecting parameters that the surrogate model predicts will yield high satisfaction, and exploration, which means selecting parameters in regions of high uncertainty to improve the model accuracy [18]. We implement a pairwise comparison querying mechanism. Instead of asking the user to rate a single audio sample on an absolute scale, which is known to be cognitively difficult and susceptible to anchoring bias, the system presents two slightly different audio samples and asks the user to indicate which one they prefer. The user feedback is mathematically treated as a binary observation, indicating that the preference score of the chosen sample is higher than that of the rejected sample. This binary observation is used to update the posterior distribution of the Gaussian process. By continuously presenting strategically selected pairs and updating the surrogate model, the system rapidly converges towards the region of the parameter space that aligns with the optimal aesthetic preference of the user. This strategy dramatically reduces the number of interactions required compared to manual slider adjustments or random search.

4. Implementation and System Architecture

4.1 System Components and Interaction Flow

The realization of the conceptual framework requires a robust and responsive system architecture capable of handling intensive audio processing tasks while maintaining a fluid user experience. The implemented architecture is divided into three primary modules encompassing the user interface front end, the active learning controller, and the neural synthesis back end. The front end is a web based application designed for intuitive interaction. It provides visual representations of the musical score and lyrics, along with a

dedicated elicitation workspace. In this workspace, users are presented with playback controls for the paired audio samples generated during the querying phase. The interface is intentionally minimalist, hiding the complex acoustic parameters to force the user to rely entirely on auditory evaluation. Once the user makes a selection indicating their preference, this binary decision is transmitted asynchronously to the active learning controller. The active learning controller acts as the central orchestrator of the system. It houses the Gaussian process surrogate model and the acquisition function logic. Upon receiving user feedback, the controller immediately recalculates the posterior distribution and determines the next optimal pair of acoustic parameters to explore [19]. It then translates these abstract parameter values into explicit conditioning tensors required by the synthesis engine.

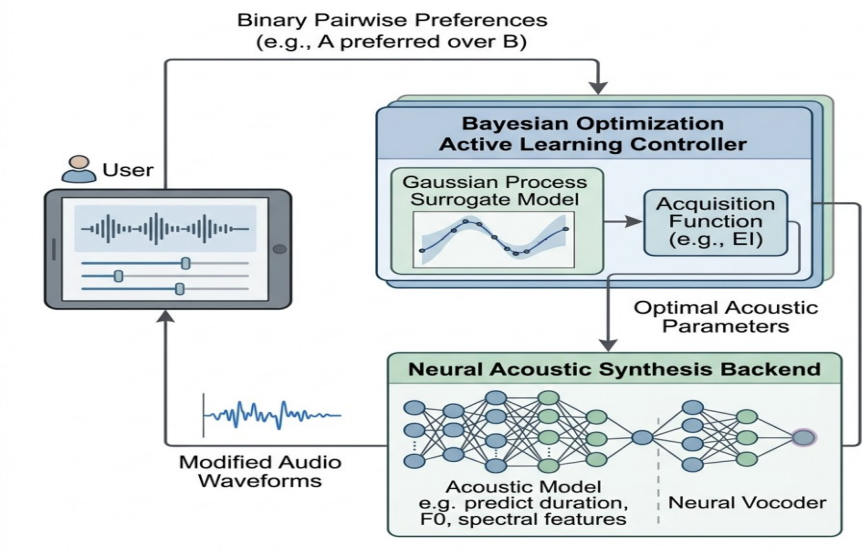

Figure 1: Comprehensive Schematic of the Interactive Singing Synthesis Architecture

4.2 Data Pipeline and Synthesis Engine Back End

The neural synthesis back end is responsible for transforming the linguistic, musical, and acoustic conditioning data into high fidelity audio waveforms. This module is built upon a modified non autoregressive text to speech architecture adapted specifically for singing. The data pipeline begins with the ingestion of the musical score, where notes are converted into precise fundamental frequency trajectories and durations are mapped to phoneme boundaries. The explicit conditioning tensors generated by the active learning controller are integrated into the network at multiple stages. The spectral energy and vocal tension parameters are concatenated with the encoder outputs before passing through the variance adaptors, allowing them to influence the overall timbral structure of the predicted mel spectrogram. The pitch dynamic parameters directly modulate the fundamental frequency contour before it is fed into the decoder. This deep integration ensures that the stylistic modifications requested by the preference model are holistically incorporated into the generated audio rather than merely layered on top as post processing effects. Once the decoder produces the optimized mel spectrogram, a high speed generative adversarial network based vocoder is employed to invert the spectrogram into a raw audio waveform. The entire synthesis pipeline is heavily optimized using hardware acceleration and parallel processing techniques to ensure that the generation of a new audio pair occurs within a strict latency threshold. This low latency is absolutely crucial for maintaining the engagement of the user during the interactive elicitation loop, as delays between providing feedback and hearing the subsequent samples can disrupt the auditory memory and degrade the quality of the preference evaluation.

5. Experimental Setup, Results, and Discussion

5.1 Experimental Design

To rigorously evaluate the effectiveness of the proposed online preference elicitation framework, a comprehensive user study was conducted. The study aimed to compare the performance of our active learning system against a traditional manual interface baseline. A diverse cohort of participants was recruited, comprising both individuals with formal audio engineering experience and individuals with no background in music production, to ensure the findings were applicable across different user skill levels [20]. The experimental task required participants to stylize a standard vocal melody to match a specific, emotionally descriptive prompt, such as creating a melancholic and intimate vocal or a powerful and

aggressive performance. Participants were divided into two experimental groups. The control group utilized a traditional dashboard interface featuring standard sliders for vibrato depth, breathiness, tension, and pitch drift. They were given a fixed time limit to manually adjust these sliders until they were satisfied with the output. The experimental group utilized the proposed online preference elicitation interface. They were presented with successive pairs of audio samples and simply had to select the one that sounded closer to the target emotional prompt. To prevent fatigue, the elicitation loop was capped at twenty iterations per task. We collected both objective system data, tracking the convergence of the parameter values over time, and subjective perceptual data through post task questionnaires.

5.2 Subjective Evaluation Metrics

The subjective evaluation focused on three primary dimensions encompassing naturalness, similarity to intention, and perceived cognitive load. Naturalness was measured to ensure that the parameter manipulations induced by the preference model did not introduce unnatural artifacts or degrade the baseline quality of the neural vocoder. Similarity to intention measured how closely the final synthesized audio aligned with the internal goal of the user as dictated by the emotional prompt [21]. Cognitive load assessed the mental effort required to complete the task using the assigned interface. Responses were recorded using a standard Mean Opinion Score scale ranging from one to five.

Table 2: Mean Opinion Score Results for Different Interfaces

Evaluation Metric	Control Group Manual Interface	Experimental Group Online Elicitation
Naturalness	4.12	4.25
Similarity to Intention	3.45	4.60
Perceived Control	3.80	4.15
Ease of Use	2.55	4.75

The results presented in the table clearly demonstrate the superiority of the online elicitation approach across almost all subjective metrics. While the naturalness scores were statistically similar, indicating that both methods maintained high audio fidelity, the similarity to intention score was significantly higher for the experimental group. This suggests that users were much more successful at achieving their desired aesthetic outcome when guided by the active learning system than when attempting to manually manipulate the acoustic space. The most striking difference was observed in the ease of use and perceived control metrics. Despite having direct access to every parameter, the control group reported lower perceived control, reflecting the difficulty of translating an acoustic concept into specific slider values. Conversely, the experimental group found the pairwise comparison process highly intuitive and effortless.

5.3 Objective Convergence Analysis

In addition to subjective feedback, we analyzed the objective convergence behavior of the Gaussian process surrogate model during the elicitation sessions. For each participant in the experimental group, we tracked the Euclidean distance between the parameter vector selected at each iteration and the final optimized parameter vector reached at the end of the session. The analysis revealed a rapid initial decrease in distance, indicating that the active learning algorithm was highly efficient at eliminating unfavorable regions of the parameter space within the first few queries. On average, the model reached a stable region characterized by minor localized exploration after approximately twelve iterations. This rapid convergence is critical for the practical viability of the system, as it proves that meaningful personalization can be achieved in a matter of minutes without subjecting the user to prolonged and tedious evaluation periods. We also observed distinct clustering in the final optimized parameter spaces based on the emotional prompts. For example, prompts requiring an intimate vocal consistently converged towards parameter vectors characterized by high breathiness, delayed vibrato, and lower vocal tension, confirming that the parameterization accurately captured the perceptual intent.

5.4 Discussion of Findings

The empirical results strongly support the hypothesis that online preference elicitation is a more effective and user friendly paradigm for personalized audio generation than traditional manual parameterization. The success of the system can be attributed to several factors. First, human auditory memory is highly comparative. We are much better at determining which of two sounds is more pleasing than we are at assigning an absolute numerical value to a single sound characteristic [22]. The pairwise comparison strategy leverages this cognitive strength. Second, the Bayesian optimization approach successfully manages the complex, non linear interactions between different acoustic parameters. Adjusting breathiness often changes the perception of vibrato, making manual adjustment a frustrating process of trial and error. The active learning model implicitly learns these dependencies and navigates the multidimensional space holistically. Furthermore, the system democratizes advanced synthesis technology. By abstracting the complex signal processing parameters behind a simple preference interface, creators without technical

backgrounds can achieve highly nuanced and professional level vocal stylizations. However, the study also revealed certain challenges. The reliance on pairwise comparisons means the user cannot directly jump to a known configuration if they already possess the technical knowledge to do so. Therefore, an optimal commercial system might require a hybrid approach, offering the elicitation loop for exploration and discovery while maintaining an advanced mode for precise, manual micro adjustments.

6. Conclusion and Future Directions

6.1 Summary of Contributions

This paper has introduced a novel and comprehensive framework for integrating online preference elicitation into personalized singing synthesis. We have addressed the critical limitation of static, generalized audio generation by proposing a human in the loop methodology that dynamically adapts the acoustic output to the subjective aesthetic tastes of individual users. By conceptualizing a robust parameterization of vocal characteristics and implementing a Bayesian optimization based active learning strategy, we created an intuitive interactive system that bypasses the need for complex manual slider interfaces. The rigorous experimental evaluation provided compelling evidence that our proposed system significantly outperforms traditional methods in terms of user satisfaction, ease of use, and the ability to accurately capture complex emotional intentions. The objective convergence analysis further proved the efficiency of the active elicitation algorithm, demonstrating that optimal personalization can be achieved with minimal cognitive burden on the user. This research represents a significant step forward in making advanced generative artificial intelligence more accessible and responsive to human creativity.

6.2 Limitations and Future Work

Despite the promising results, several limitations within the current framework present opportunities for future research. One primary limitation is the inherent temporal cost of auditory evaluation. While the active learning algorithm minimizes the number of queries, listening to full audio phrases still requires time. Future iterations could explore predictive truncation techniques, where the system stops playing a sample as soon as it detects that the user has formed a definitive preference, further accelerating the elicitation loop. Another limitation is the reliance on explicit binary feedback. Expanding the elicitation mechanism to incorporate multi modal implicit feedback, such as tracking user facial expressions, pupillary response, or even neurological signals during the listening phase, could provide a richer and less intrusive estimation of user preference. Furthermore, the current system personalizes the acoustic parameters for a specific phrase or song. Future work should investigate the longitudinal modeling of user preferences to create persistent digital vocal personas that retain the stylized characteristics of a user across different musical contexts and sessions. Finally, expanding the parameter space to include broader stylistic variations, such as incorporating genre specific vocal embellishments like growls, falsetto flips, or rhythmic phrasing adjustments, will further enhance the expressive capability and personalization depth of the singing synthesis engine. Integrating these advancements will continue to blur the line between technical synthesis generation and intuitive artistic expression.

References

1. 1. 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.
2. 2. 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).
3. 3. 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.
4. 4. 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.
5. 5. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of ICML.
6. 6. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of ICML.
7. 7. 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.
8. 8. 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.
9. 9. Sun, L., Hu, J., Li, M., & Peng, H. (2024, July). R-ode: Ricci curvature tells when you will be informed. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2594-2598).
10. 10. Gu, C., Zhang, W., Huang, Z., Kou, J., Liu, Z., Zhao, C., Liu, C., Zhang, L., Lin, W., Wang, Z., Deng, J., Xie, Y., Huang, G., Zhang, C., Lu, X., Wang, C., Zhang, Z., Yuan, H., Duan, X., & Fang, Y. (2024). LENS: Layers of evaluation of hallucination in GenAI systems. In 2024 7th International Conference on Universal Village (UV) (pp. 1–85). IEEE. https://doi.org/10.1109/UV63228.2024.11189150

11. 11. Yao, Y., Zhu, Z., Miao, P., Cheng, X., Shu, F., & Wang, J. (2025). Optimizing hybrid RIS-aided ISAC systems in V2X networks: A deep reinforcement learning method for anti-eavesdropping techniques. IEEE Transactions on Vehicular Technology, 74 (6), 9224-9239.
12. 12. Wei, M., Li, R., Wang, L., Chang, Z., Xu, L., & Han, Z. (2026). Toward Adaptive Tracking and Communication via an Airborne Maneuverable Bi-Static ISAC System. IEEE Transactions on Vehicular Technology.
13. 13. Wang, Y., He, S., Chen, G., Chen, Y., & Jiang, D. (2022, December). XLM-D: Decorate cross-lingual pre-training model as non-autoregressive neural machine translation. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 6934-6946).
14. 14. 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.
15. 15. Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.
16. 16. Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.
17. 17. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
18. 18. Zhang, Q., Yang, M., Sun, G., Xiang, Y., Wang, S., Zhang, J., & Li, S. (2026). Representation learning accelerates the development of models for Li-ion battery health diagnostics and prognostics. Energy Storage Materials, 104897.
19. 19. Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).
20. 20. Wang, Jiacheng, et al. "A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting." (2026).
21. 21. Cao, X., Tao, J., Liu, Z., Lyu, R., & Li, J. (2026). Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics. Future-Adaptive Intelligence and Lifelong Systems, 1 (1).
22. 22. Abu Saleh, A., Siouris, S., Hughes, K., Yuan, R., & Pourkashanian, M. (2023). Assessment of mixtures of iso-pentanol and Jet A-1 for use in aviation gas turbine engines. In AIAA SCITECH 2023 Forum (p. 2341).