Emotion-Aware Singing Voice Generation Through Multimodal Music and Vocal Feature Analysis

Brady Gibson
School of Computing, Clemson University, Clemson, SC, USA.
brady.gibson@clemson.edu

Abstract

Emotion-aware singing voice generation requires a model to coordinate musical score, lyric content, vocal timbre, pitch behavior, and expressive intention without collapsing these factors into a single undifferentiated style signal. Many neural singing and speech synthesis systems treat emotion as a categorical label or a global embedding, which can obscure the separate roles of prosody, spectral envelope, vibrato, breathiness, and timing. This paper develops a methodological framework for emotion-aware singing voice generation through multimodal music and vocal feature analysis. The central contribution is a factorized generative formulation in which score and lyric information define structural constraints, singer identity defines timbre constraints, and continuous valence-arousal affect modulates vocal expression through an emotion-gated fusion pathway. The framework combines frame-level vocal descriptors, including log fundamental frequency, pitch derivatives, energy, harmonic-to-noise balance, aperiodicity, and vibrato strength, with score and lyric encoders and a cross-attention mechanism for alignment-aware conditioning. A conditional generative objective is specified, together with reconstruction, adversarial, emotion-classification, and regularization terms. The analysis identifies principal failure modes, including score-audio misalignment, emotion leakage into singer identity, over-smoothed vibrato, and biased emotion supervision. It also clarifies why emotion conditioning should act separately on prosodic and timbral pathways and why evaluation must combine signal-level, emotion-level, and listener-level criteria. The resulting framework provides a technically grounded basis for expressive singing generation and for disciplined assessment of multimodal affective control.

Keywords

singing voice synthesis, emotion-aware generation, multimodal feature fusion, vocal expression, cross-attention, generative modeling

1. Introduction

Singing voice generation sits at the intersection of music information processing, speech synthesis, and affective computing. Unlike ordinary speech synthesis, singing requires precise control of pitch, duration, timing, lyric intelligibility, vocal timbre, and expressive variation. A singer does not merely pronounce lyrics to a melody; the voice carries continuous expressive information through microtonal pitch movement, vibrato, spectral tilt, breathiness, consonant timing, and dynamic emphasis. Emotion-aware singing generation therefore cannot be reduced to attaching a categorical emotion label to a conventional text-to-speech pipeline. The model must preserve the musical structure while allowing expressive variation that is perceptually coherent with the intended affect.

Deep generative models provide a natural foundation for this problem. Neural generative approaches have advanced substantially through autoregressive, variational, adversarial, and diffusion-based formulations [1]. Variational autoencoders offer latent-variable formulations

for complex acoustic distributions [2], while diffusion models provide iterative denoising objectives with strong distributional modeling capacity [3]. For audio, autoregressive waveform models demonstrated the value of high-resolution temporal modeling [4], and sequence-to-sequence speech synthesis established the feasibility of end-to-end acoustic generation from symbolic inputs [5]. Feed-forward and duration-controlled architectures improved controllability and alignment [6], while adversarial vocoders improved the fidelity of generated waveforms [7]. These developments supply the generative machinery for singing voice generation, but they do not by themselves resolve the specifically musical and affective problems of singing.

The missing link is a representation that treats emotion as a structured modulator of vocal expression rather than as a single global attribute. In singing, affect is distributed across at least two interacting pathways. The prosodic pathway controls pitch contour, timing, vibrato, and intensity. The timbral pathway controls spectral envelope, harmonic-to-noise balance, breathiness, and resonance. A model that concatenates an emotion vector with a score representation may entangle these pathways and produce either weak expressiveness or unstable pitch and timbre. The analytical position developed here is that emotion-aware singing generation benefits from a factorized conditioning scheme in which score and lyric tokens define structural constraints, singer identity defines stable timbral constraints, and continuous affect modulates selected vocal pathways through gate-controlled fusion.

This paper develops that position as a complete methodological framework. The contribution is not an empirical benchmark but a technically specified formulation for multimodal singing voice generation that integrates music score, lyric, vocal feature, singer identity, and continuous emotion representations. The framework defines a conditional generative objective, a factorized vocal feature representation, an emotion-gated fusion block, a cross-attention alignment mechanism, and a multi-term training objective. It also examines boundary conditions and failure mechanisms that are specific to singing, including score-audio misalignment, emotion leakage, over-smoothing of vibrato, and supervision bias. The discussion is intended to support subsequent implementation and evaluation without presupposing unavailable experimental results.

2. Related Work and Vocal-Music Feature Analysis

Singing voice analysis depends on representations that capture both musical structure and voice quality. Classical parametric analysis tools such as WORLD separate fundamental frequency, spectral envelope, and aperiodicity, enabling high-quality analysis, modification, and synthesis [8]. Voice-analysis toolkits such as COVAREP provide complementary descriptors of glottal source, prosody, and spectral characteristics [9]. Pitch estimation methods based on convolutional representations improve the robustness of fundamental-frequency tracking in singing and speech [10]. These tools are not merely preprocessing utilities. They define the observable variables through which emotion and musical expression become measurable. If the representation omits vibrato or harmonic-to-noise balance, for example, the model has little basis for learning the difference between a tense, pressed vocal tone and a relaxed, breathy one.

Emotion representation has its own literature and constraints. Physiological and multimedia datasets have popularized dimensional affect representations, including valence and arousal, as continuous alternatives to discrete emotion categories [11]. Music emotion recognition research has similarly emphasized the need to model continuous affect and listener-dependent variability rather than assuming a single objective label per excerpt [12].

Multimodal emotional speech and song corpora provide controlled vocal expressions that can support training and evaluation of affective models [13]. Surveys of speech emotion recognition identify features, classifiers, and database limitations that remain relevant to singing, especially the sensitivity of emotion classifiers to speaker, language, and recording conditions [14]. Open-source feature extraction frameworks have made standardized acoustic descriptors widely available [15], while research on vocal communication of emotion has emphasized that affect is conveyed through multiple acoustic cues rather than a single parameter [16].

Singing introduces additional data and modeling demands. A recent global multi-technique singing corpus provides realistic music scores and annotations intended to support a broad range of singing tasks [17]. Such resources are relevant because emotion-aware generation requires alignment among symbolic score, lyric content, and acoustic realization. Style-control research in neural speech synthesis has shown that unsupervised style representations can capture global and local variation beyond explicit labels [18]. For singing, however, style and emotion are not identical. Style may include singer-specific habits, genre conventions, and recording aesthetics, whereas emotion concerns the expressive intention and perceived affect of a performance. A usable framework must preserve this distinction.

The gap addressed here is therefore representational and architectural. Existing analysis tools provide strong low-level vocal descriptors; existing emotion research provides dimensional and categorical affect constructs; and neural generative models provide flexible acoustic decoders. What remains under-specified is how to combine these elements so that emotion conditions singing expression without destabilizing score alignment or singer identity. The framework below responds to that gap by separating structural, timbral, and affective factors and by making their interaction explicit.

3. Problem Formulation and Multimodal Representation

Let x denote a singing waveform or a sequence of acoustic frames, s a musical score representation containing pitch, duration, phrasing, and dynamic marks, l a lyric or phoneme sequence, e a continuous affect vector, and b a singer-identity representation. The generation target is the conditional distribution of singing audio given score, lyric, affect, and singer constraints. Because the acoustic realization involves latent variation, the model introduces a latent variable z that captures residual factors such as microtiming, articulatory detail, and recording-dependent variation. The conditional generation problem is written as an intractable marginal over z.

$$p(x \mid s, l, e, b) = \int p(x \mid z, s, l, e, b)\, p(z \mid s, l, e, b)\, dz$$

Here p(z | s, l, e, b) is the conditional prior over latent variation, and p(x | z, s, l, e, b) is the acoustic decoder. The integral indicates that exact likelihood evaluation is generally unavailable, motivating amortized variational inference, adversarial training, or diffusion-based denoising. The formulation is deliberately factorized: score and lyric variables constrain what must be sung, b constrains who is singing, and e constrains how the vocal expression is shaped. This separation matters because a model that treats all conditioning as a single vector cannot easily diagnose whether an expressive variation comes from emotion, singer identity, or score-driven necessity.

Affect is represented in a continuous two-dimensional valence-arousal space, with an optional intensity derived from the vector magnitude. Valence indicates pleasant versus unpleasant affect, and arousal indicates calm versus activated affect. Intensity separates

mildly expressive from strongly expressive renditions that occupy similar regions of the valence-arousal plane.

$$e = [v, a]\text{^}T \in [-1, 1]\text{^}2, \rho = \sqrt{(v\text{^}2 + a\text{^}2)}$$

In this expression, v is valence, a is arousal, and ρ is affect intensity. The squared norm is not treated as a complete emotion descriptor, because two performances can share intensity while differing in valence or arousal. Its role is to provide a scalar control that can be separated from directional affect. This distinction is useful in singing because increasing emotional intensity may raise vibrato extent, dynamic contrast, and spectral brightness without changing the qualitative direction of the emotion.

For frame-level acoustic analysis, the vocal representation combines pitch, energy, and voice-quality descriptors. At frame index t, the feature vector includes log fundamental frequency, its local derivative, energy, harmonic-to-noise balance, aperiodicity, and a vibrato-strength descriptor. The derivative captures pitch movement, while the voice-quality terms capture roughness, breathiness, and periodicity.

$$v_t = [\log f0_t, \Delta \log f0_t, E_t, H_t, A_t, V_t]^T$$

The term f0_t is fundamental frequency, Δ log f0_t is its frame-to-frame change, E_t is frame energy, H_t is harmonic-to-noise balance, A_t is aperiodicity, and V_t is a local vibrato-strength estimate. These variables are not interchangeable with emotion labels. They are intermediate acoustic evidence from which expressive variation can be generated and evaluated. For example, high arousal may correlate with greater energy and wider pitch movement, but the same acoustic pattern may appear in a musically loud passage without a strongly emotional intent. The model therefore uses e as a conditioning variable and v_t as observable vocal evidence, rather than assuming a deterministic mapping between them.

4. Emotion-Aware Generation Framework

The proposed architecture contains four encoders and a conditioned decoder. A score encoder maps symbolic pitch, duration, and phrasing into a temporal representation. A lyric encoder maps phoneme or character sequences into linguistic embeddings. A vocal-feature encoder summarizes observed or target acoustic descriptors such as those in v_t. An emotion encoder maps the valence-arousal vector e and its intensity into an affective embedding. Singer identity b is encoded separately and injected into the decoder through normalization or embedding layers. The separation of b from e is important because emotion should alter expression without erasing singer-specific timbre.

The central fusion mechanism is emotion-gated modulation. Rather than concatenating e with the score representation and relying on the decoder to discover useful interactions, the model computes a time-varying gate that controls how much vocal-feature information and score information contribute to the fused state. The gate is conditioned jointly on affect and on the local score context. This design allows emotion to change the balance between stable musical constraints and expressive vocal variation at different moments in a phrase.

$$\alpha_t = \sigma(W_e\, e + W_s\, s_t + b), h_t = \alpha_t \odot h_t\text{^}v + (1 - \alpha_t) \odot h_t\text{^}s$$

Here σ is the logistic sigmoid, W_e and W_s are learned projection matrices, b is a bias vector, s_t is the local score representation, h_t^v is the vocal-feature representation, and h_t^s is the score representation. The symbol ⊙ denotes elementwise multiplication. When α_t is large, the fused state emphasizes vocal expression; when it is small, the state emphasizes score-derived structure. The gate is not a global emotion switch. It varies with

time and score context, which matters for singing because expressive deviation must remain coordinated with melodic and rhythmic constraints.

Alignment between symbolic structure and acoustic realization is handled through cross-attention. The vocal or acoustic query attends to score, lyric, and emotion key-value representations, producing a context vector for each output frame. This mechanism supports soft alignment while allowing the decoder to access long-range musical context.

$$C = softmax(Q\,K\text{^}T / \sqrt{d})\,V$$

In this expression, Q is the query matrix derived from the current acoustic or vocal state, K and V are key and value matrices derived from score, lyric, and affective embeddings, and d is the query-key dimension used for scaling. C is the resulting context matrix. Cross-attention does not by itself guarantee correct timing; it must be combined with duration supervision, monotonic alignment constraints, or score-derived positional features. For singing, an unconstrained attention mechanism can smear consonant onsets or shift sustained notes. The framework therefore treats alignment as a joint requirement of attention, duration modeling, and score-position encoding.

The decoder generates acoustic frames or waveform samples conditioned on the fused representation, the singer embedding, and the latent variable z. Any suitable conditional generative backbone can be used, including autoregressive, adversarial, or diffusion-based decoders. The choice affects sample quality, inference cost, and controllability rather than the conceptual structure of conditioning. The key requirement is that the decoder expose separate injection points for structural, timbral, and affective information. This separation supports diagnostic control: a practitioner can alter e while holding s, l, and b fixed, or alter b while holding e fixed, and then observe whether the intended factor changes without unintended changes in the others.

5. Training Objectives and Inference

Training requires the model to reconstruct acoustic detail, produce perceptually plausible singing, respect the intended emotion, and avoid unstable or unnatural variation. These requirements are expressed through a weighted multi-term objective. The reconstruction term compares generated acoustic features or spectrograms with reference singing under multiple resolutions. The adversarial term encourages waveform realism. The emotion term aligns generated expressive characteristics with the intended affect. The regularization term controls smoothness, pitch continuity, and latent-scale behavior.

$$L = \lambda_r ecL_r ec + \lambda_a dvL_a dv + \lambda_e moL_e mo + \lambda_r egL_r eg$$

The coefficients λ_rec, λ_adv, λ_emo, and λ_reg are nonnegative weights that set the relative importance of reconstruction, adversarial realism, affective consistency, and regularization. L_rec may combine frame-level feature distance with multi-resolution spectral loss so that both fine temporal detail and broader spectral structure are considered. L_adv may use a Wasserstein formulation with gradient penalty to stabilize adversarial training [19], following the improved training procedure for Wasserstein generative adversarial networks [20]. L_reg may penalize excessive frame-to-frame pitch discontinuities, unstable vibrato-rate changes, and latent outliers.

The emotion-consistency term uses an auxiliary emotion classifier or regressor applied to generated audio features. In the categorical form, the classifier outputs a distribution over K

affect regions, while the continuous valence-arousal vector remains the primary conditioning input. This hybrid design allows stable supervision without discarding continuous control.

$$L_emo = -\sum_k y_k \log q_k(\hat{x})$$

Here k indexes affect regions, y_k is the target distribution or one-hot label for the intended region, q_k(x̂) is the classifier probability assigned to generated singing x̂, and the sum is over K regions. The term is auxiliary rather than definitional: it regularizes generated expression toward the intended affect but does not prove that a listener will perceive that affect. This distinction is important because emotion classifiers can exploit singer identity, recording quality, or lexical cues. A model that minimizes L_emo too aggressively may produce stereotyped emotional markers rather than musically coherent expression.

Inference can be performed deterministically by using the prior mean of z, or stochastically by sampling z and averaging or selecting among generated candidates. The affect vector e can be supplied directly as a continuous control or inferred from a reference performance. When e is inferred, the system should expose a confidence estimate and should avoid forcing strong affect changes when the posterior is ambiguous. A practical inference rule is to apply affect modulation only above a calibrated confidence threshold and to limit the rate of change of expressive parameters across adjacent notes. This rule reduces abrupt shifts in vibrato, dynamics, or timbre that can sound unnatural in sustained singing.

6. Technical Discussion and Boundary Conditions

The framework's main technical claim is that emotion-aware singing generation benefits from factorized conditioning and pathway-specific modulation. This claim has several consequences. First, score-audio alignment must be treated as a first-class constraint. If the decoder attends too freely to lyric or emotion representations, it may produce expressive vowel coloration at the expense of correct phoneme timing. Duration supervision, score-position features, and monotonic alignment constraints are therefore not optional refinements; they are necessary to prevent expressive modulation from disrupting intelligibility.

Second, emotion conditioning must be disentangled from singer identity. Singer embeddings can absorb affect-correlated recording characteristics, such as breathiness, microphone proximity, or vocal effort. If the same singer appears mainly in high-arousal recordings, the model may confuse that singer's identity with arousal. Counterfactual training pairs, singer-balanced affect sampling, and adversarial identity regularization can reduce this leakage. The objective should reward changes in expression when e changes while penalizing changes in singer-discriminative features when only e is varied.

Third, vibrato and pitch modulation require explicit treatment. Over-smoothed pitch contours are a common failure mode in neural acoustic models because frame-level regression tends to average periodic variation. Emotion-aware singing cannot afford to erase vibrato, because vibrato rate and extent are perceptually salient expressive cues. Feature losses on pitch derivatives, energy contours, and local modulation statistics can preserve these structures better than frame-level spectral loss alone. At the same time, excessive emphasis on vibrato can produce mechanical oscillation, so regularization must distinguish natural quasi-periodic modulation from unstable jitter.

Fourth, emotion supervision is vulnerable to label bias. Discrete emotion labels compress culturally variable and listener-dependent interpretations into fixed categories. Continuous valence-arousal annotations reduce but do not eliminate this problem. A model trained on a

narrow set of singers, genres, or languages may learn affect markers that do not generalize. Evaluation should therefore include held-out singers, cross-genre material, and listener ratings rather than relying exclusively on an automatic emotion classifier. The singing corpus discussed above [17] is relevant because multi-technique and score-aligned data can support such controls, but no single corpus removes the need for explicit generalization testing.

The boundary conditions of the framework are also clear. It assumes access to a musical score or a reliable symbolic proxy, a lyric or phoneme sequence, and a singer identity representation. It is most directly applicable to monophonic or predominantly solo singing with manageable accompaniment. Polyphonic mixtures require source separation, score alignment, and accompaniment-aware conditioning before the same formulation can be applied. Cross-lingual singing introduces phoneme inventory and prosody mismatches that may require language-specific encoders or adaptation. Real-time applications impose latency constraints that may favor feed-forward or distilled decoders over iterative diffusion sampling.

Evaluation should accordingly be multi-level. At the signal level, pitch accuracy, spectral distance, energy contour fidelity, and vibrato statistics measure whether musical structure and vocal quality are preserved. At the emotion level, classification or regression against held-out affect annotations measures whether intended affect is encoded, with the caveat that classifier accuracy is not equivalent to perceived expressiveness. At the listener level, perceptual tests should assess naturalness, emotional appropriateness, singer consistency, and lyric intelligibility. Ablation studies should isolate the effects of emotion gating, cross-attention alignment, and vocal-feature losses. These evaluations are not reported here as empirical findings; they define the evidential standard that a full implementation would need to meet.

7. Conclusion

Emotion-aware singing voice generation requires more than adding an affect label to a neural synthesizer. The score, lyric, singer identity, vocal quality, and expressive intention interact at multiple timescales, and a model must preserve musical structure while allowing controlled affective variation. This paper has developed a methodological framework that addresses this requirement through multimodal feature analysis and factorized conditioning. Continuous valence-arousal affect modulates vocal expression through an emotion-gated fusion mechanism, while cross-attention and score-derived constraints maintain alignment between symbolic structure and acoustic realization. Vocal descriptors such as log fundamental frequency, pitch derivatives, energy, harmonic-to-noise balance, aperiodicity, and vibrato strength provide intermediate variables that connect affect to measurable acoustic behavior without assuming a deterministic mapping. The training formulation combines reconstruction, adversarial, emotion-consistency, and regularization terms, and the discussion identifies alignment failure, identity leakage, vibrato smoothing, and supervision bias as central technical risks. The framework's practical value lies in making these interactions explicit and testable. It supports the design of singing generation systems that can vary emotional expression while retaining singer consistency and musical control, and it provides a basis for evaluation that treats acoustic fidelity, affective intent, and listener perception as distinct but connected criteria.

References

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

2. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations.

3. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.

4. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

5. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech (pp. 4006–4010).

6. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems, 32, 3165–3174.

7. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33, 17022–17033.

8. Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A high-quality speech analysis, modification and synthesis system. Acoustical Science and Technology, 37(1), 33–39.

9. Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP: A collaborative voice analysis repository for speech technologies. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 960–964).

10. Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). CREPE: A convolutional representation for pitch estimation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 161–165).

11. Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18–31.

12. Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology, 3(3), Article 40.

13. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.

14. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587.

15. Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia (pp. 1459–1462).

16. Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1–2), 227–256.

17. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. Advances in Neural Information Processing Systems, 37, 1117-1140.

18. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Proceedings of the 35th International Conference on Machine Learning (pp. 5180–5189).

19. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.

20. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30, 5767–5777.