Personalized Neural Voice Generation Using Multi-Attribute Speaker Embeddings and Adaptive Style Conditioning

Gioele Verri
Department of Computer Science, George Mason University, Fairfax, VA, USA.
gioelev@gmu.edu

Nurul Laksmiwati
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
contactnurul@colostate.edu

Abstract

Personalized neural voice generation must satisfy two requirements that are often in tension: preserving a target talker's identity with high fidelity and allowing expressive style to be controlled without entangling it with speaker characteristics. Conventional systems encode identity as a single speaker vector and condition synthesis on a global style label or reference utterance, which limits interpretability, compositional control, and robustness when reference speech contains channel or prosodic variability. This manuscript develops a methodological framework for personalized voice generation that represents speaker identity through multiple attribute-specific embeddings and conditions a neural synthesizer through an adaptive style pathway. The framework decomposes speaker information into timbre, pitch range, tempo, energy, articulation, and channel-related subspaces; combines them through a learned reliability gate; and uses a style encoder that infers continuous style latents from reference speech. Style and speaker information are injected into the decoder through adaptive normalization, cross-attention, and guidance, with explicit regularization to reduce identity leakage. A training objective combines reconstruction, speaker classification, style modeling, adversarial discrimination, Kullback-Leibler regularization, and orthogonality constraints. The analytical assessment explains why attribute-gated embeddings improve interpretability and partial personalization, how adaptive conditioning supports style interpolation and classifier-free guidance, and which failure modes arise from attribute entanglement, short reference utterances, channel bias, and over-regularization. The resulting framework offers a technically grounded design for controllable personalized speech synthesis and clarifies the conditions under which multi-attribute conditioning is preferable to monolithic speaker embeddings.

Keywords

personalized speech synthesis; multi-attribute speaker embedding; adaptive style conditioning; disentangled representation; diffusion decoder; controllable text-to-speech

1. Introduction

Neural text-to-speech has progressed from autoregressive sequence models that predict mel spectrograms [1], [2] to feed-forward architectures that reduce inference latency and improve stability [3]. Multispeaker synthesis subsequently introduced speaker-conditioned training so that a single model can generate different voices, with speaker embeddings derived

from reference utterances and verification-style objectives [4]. These developments make personalization technically feasible, but they do not by themselves resolve a central modeling problem: a compact speaker vector conflates timbre, pitch range, speaking rate, energy, articulation, and recording channel, while expressive style is often represented as a separate but weakly constrained factor. The consequence is that personalization may change the intended voice when style is adjusted, and style may drift when a target speaker is introduced.

Speaker embeddings trained for verification are effective at discriminating talkers [5], and channel-attentive architectures such as ECAPA-TDNN improve robustness by aggregating frame-level statistics over time [6]. However, a discriminative speaker representation is optimized to answer whether two utterances belong to the same talker, not to expose which attribute of the voice should be modified. Style control has been approached through unsupervised style tokens [7] and reference prosody transfer [8], while end-to-end generative models with variational and adversarial training have improved naturalness and sampling efficiency [9]. In singing voice synthesis, style transfer and multi-level style control have been investigated as a way to separate timbre from expressive delivery [10]. These lines of work suggest that personalization and style control should be treated as interacting factors rather than independent conditioning inputs.

This manuscript proposes a methodological framework for personalized neural voice generation using multi-attribute speaker embeddings and adaptive style conditioning. The central contribution is a conditioning architecture in which speaker identity is represented by attribute-specific embeddings with learned reliability weights, and style is represented by a continuous latent inferred from reference speech and injected into the decoder through adaptive normalization and gated cross-attention. The framework does not assume that attribute labels are fully observed at inference; it supports partial personalization when only some attributes are specified and falls back to a learned prior when reference style is unavailable. The design is analyzed with respect to representation capacity, disentanglement, conditioning stability, and inference-time control.

2. Related Work and Problem Formulation

2.1 Generative backbones and conditional speech synthesis

Diffusion and score-based decoders model speech distributions through iterative denoising and have been applied to text-to-speech with mel-spectrogram or waveform generation [11], [12]. Such decoders are attractive for personalized generation because conditioning can be injected at multiple denoising steps, allowing speaker and style information to influence coarse and fine acoustic structure. Zero-shot multi-speaker synthesis has been explored with speaker-conditioned models that generalize from short reference utterances [13], and large-scale in-context speech generation has demonstrated the value of broad data coverage and flexible prompting [14]. Feature-wise linear modulation provides a general mechanism for injecting conditioning signals into intermediate feature maps [15], while denoising diffusion probabilistic models provide the formal generative objective used by many recent speech systems [16]. Discrete latent representations offer an alternative route for compression and token-level modeling [17], and latent diffusion shows how generation can be performed in a lower-dimensional space rather than directly in waveform or high-resolution spectral domains [18].

2.2 Style conditioning and speaker-style interaction

Style tokens provide unsupervised control over prosodic variation and can be transferred across speakers [7], while reference-based prosody transfer conditions synthesis on temporal prosodic patterns from an utterance [8]. These methods improve expressiveness but do not guarantee that speaker identity remains invariant under style manipulation. Adaptive normalization addresses this issue by generating scale and bias parameters from a conditioning vector, so that style can modulate shared feature statistics without replacing the speaker embedding. The risk is that the conditioning network may learn to encode speaker information in the same modulation parameters, producing identity leakage. The proposed framework therefore separates speaker and style pathways and applies orthogonalization and adversarial constraints to discourage the style encoder from representing talker identity.

2.3 Formal problem statement

Let x_text denote a text sequence, a = {a_1, ..., a_M} a set of speaker-attribute observations, y_ref a reference style utterance, and ŷ the generated acoustic representation or waveform. The synthesis task is to learn a generator that maps text, a multi-attribute speaker representation, and a style latent to an output consistent with the requested identity and style. This relationship is written as

$$\hat{y} = G\theta(x_text, e_spk, z_style)$$

where Gθ is the synthesis network, e_spk is the multi-attribute speaker embedding, z_style is the style latent, and θ denotes trainable parameters. The formulation makes explicit that personalization and style are separate conditioning inputs at the interface, even though they interact inside the decoder. A useful system must preserve e_spk when z_style changes, and it must allow z_style to vary when e_spk is fixed.

3. Multi-Attribute Speaker Representation

3.1 Attribute-specific encoders and reliability gating

The speaker representation is constructed from attribute-specific encoders rather than a single pooled vector. For each attribute m, an encoder E_m maps an observation a_m, such as a short reference utterance, a pitch-range estimate, or a channel descriptor, to an embedding e_m. A reliability gate assigns weights according to the availability and estimated quality of each attribute, so that a noisy pitch estimate does not dominate a clean timbre estimate. The combined speaker embedding is

$$e_spk = \Sigma_m \alpha_m E_m(a_m), \Sigma_m \alpha_m = 1$$

where α_m is the normalized gate weight for attribute m and e_m = E_m(a_m). The gate is produced by a small feed-forward network that receives attribute embeddings and confidence features. This design supports partial personalization: if tempo is unspecified, its weight can be reduced and the remaining attributes can still define a coherent voice. It also improves interpretability, because changing the pitch-range embedding should alter pitch statistics more directly than changing a monolithic speaker code.

The attributes should not be treated as fully independent physical quantities. Timbre, pitch range, and articulation covary across talkers, and channel characteristics can mimic timbre differences. The model therefore uses attribute dropout during training, randomly masking one or more a_m and forcing the gate to rely on the remaining evidence. Attribute dropout improves robustness to missing attributes and reduces the chance that the decoder treats a single attribute as a proxy for the entire speaker identity.

3.2 Disentangling speaker and style

A multi-attribute embedding is useful only if the style pathway does not absorb speaker information. The style encoder is trained on reference acoustic features with a variational or diffusion objective, while the speaker encoder is trained with a speaker classification loss. An adversarial style classifier is applied to z_style to discourage predictable speaker identity, and an orthogonality penalty reduces correlation between the speaker and style embedding spaces. These constraints are complementary: adversarial training suppresses linearly decodable identity, while orthogonality provides a stable gradient signal when the adversarial classifier is weak or unstable.

The orthogonality penalty is defined over centered embedding matrices E_spk and E_style formed from a batch of utterances. It penalizes the squared Frobenius norm of their cross-covariance, which encourages the two subspaces to vary independently. This does not prove semantic disentanglement, but it reduces a common failure mode in which style latents encode recording channel or vocal-tract characteristics that are also used for speaker identification.

4. Adaptive Style Conditioning

4.1 Style inference from reference speech

The style encoder maps a reference utterance y_ref and, optionally, the speaker embedding e_spk to a latent distribution. Conditioning the posterior on e_spk allows the encoder to separate style from the speaker's baseline voice, but it also creates a path for identity leakage. The posterior is therefore regularized toward a prior that depends on text and speaker identity but not on detailed prosodic timing. A diagonal Gaussian posterior is used for exposition:

$$q\psi(z_s \mid y_ref, e_spk) = N(\mu\psi(y_ref, e_spk), diag(\sigma\psi^2(y_ref, e_spk)))$$

where μψ and σψ are mean and standard-deviation outputs of the style encoder, ψ denotes encoder parameters, and z_s is the style latent. During training, z_s is sampled from this posterior; during inference, the mean is used when deterministic style transfer is desired, while sampling is used when stylistic variation is requested. The prior p(z_s | x_text, e_spk) supplies a fallback when no reference utterance is available and prevents the posterior from becoming an unconstrained copy of the reference acoustics.

4.2 Adaptive normalization and gated cross-attention

The decoder receives z_s and e_spk through two complementary mechanisms. Adaptive normalization generates per-layer scale and bias parameters from the conditioning vector, which is effective for global style properties such as energy, tempo, and pitch range. Gated cross-attention allows local style patterns, such as phrase-final lengthening or emphasis placement, to influence the decoder at specific time steps. For layer l, the normalization operation is

$$h_l = \gamma_l(z_s, e_spk) \odot Norm(h_\{l-1\}) + \beta_l(z_s, e_spk)$$

where h_{l-1} is the input feature map, Norm denotes layer or instance normalization, γ_l and β_l are conditioning functions, and ⊙ is elementwise multiplication. The functions γ_l and β_l are produced by a small modulation network that receives the concatenation of z_s and e_spk. Separating the modulation network from the speaker encoder prevents the speaker embedding from being overwritten by style inference.

Gated cross-attention is used in parallel with adaptive normalization. Style features serve as keys and values, decoder states serve as queries, and a learned gate controls how much local style information is admitted at each layer. The gate is conditioned on both z_s and e_spk, so the model can reduce style attention when the reference utterance is short, noisy, or mismatched to the target speaker. This gating mechanism is important in personalized generation because style transfer from a different talker can introduce incompatible prosodic patterns.

4.3 Classifier-free guidance and style interpolation

Classifier-free guidance provides a practical way to trade off adherence to the conditioning signal against sample diversity. The denoising network predicts noise under the full condition c = (x_text, e_spk, z_s) and under a null condition $\varnothing$. The guided prediction is

$$\hat{\varepsilon} = \varepsilon\theta(x_t, t, \emptyset) + s\,(\varepsilon\theta(x_t, t, c) - \varepsilon\theta(x_t, t, \emptyset))$$

where εθ is the noise-prediction network, x_t is the noisy acoustic variable at step t, s is the guidance scale, and $\hat{\varepsilon}$ is the guided noise estimate. A larger s increases conditioning strength but can reduce naturalness and amplify artifacts when the condition is unreliable. The framework therefore uses a moderate guidance scale for identity and a separately tuned scale for style, allowing personalization strength and stylistic expressiveness to be adjusted independently.

Style interpolation is performed in the latent space rather than by mixing output waveforms. Given two style latents z_s1 and z_s2, an interpolated style is

$$z_s(\alpha) = normalize((1 - \alpha)\, z_s1 + \alpha\, z_s2)$$

where $\alpha \in [0, 1]$ controls the interpolation and normalize projects the result to a norm range consistent with the training prior. Latent interpolation produces smoother transitions than waveform mixing because it operates on the factors that condition the decoder. It also makes style control compositional: a user can interpolate between a neutral and an expressive reference while keeping the speaker embedding fixed.

5. Training and Inference

5.1 Composite objective

The model is trained with a composite objective that balances reconstruction quality, speaker fidelity, style modeling, adversarial separation, and latent regularization. A representative objective is

$$L = L_rec + \lambda_id L_id + \lambda_sty L_sty + \lambda_adv L_adv + \lambda_kl L_kl + \lambda_orth L_orth$$

where L_rec is the reconstruction loss between generated and target acoustic features, L_id is a speaker classification or verification loss, L_sty measures style fidelity when a reference style is available, L_adv penalizes speaker predictability from z_s, L_kl regularizes the style posterior toward its prior, and L_orth penalizes cross-covariance between speaker and style embeddings. The coefficients λ_id, λ_sty, λ_adv, λ_kl, and λ_orth control the trade-offs. Increasing λ_id improves identity preservation but may reduce style flexibility; increasing λ_adv reduces identity leakage but can destabilize training if the adversarial classifier becomes too strong. A scheduled adversarial weight and gradient reversal layer are appropriate because they prevent early training from collapsing the style encoder.

5.2 Training schedule and data conditions

The training schedule uses attribute dropout, style-reference dropout, and text perturbation. Attribute dropout forces the gate to operate with incomplete speaker descriptions. Style-reference dropout trains the prior pathway so that synthesis remains possible when no reference style is supplied. Text perturbation improves robustness to minor tokenization and phrasing differences. The model should be trained on speech that covers multiple talkers, recording conditions, speaking rates, and expressive styles; otherwise the attribute encoders learn narrow distributions and the reliability gate becomes overconfident.

The speaker and style encoders should not be trained with the same learning-rate schedule as the decoder. Speaker encoders benefit from a lower learning rate and from being frozen for short intervals, while the style encoder requires sufficient gradient flow to learn continuous prosodic variation. A two-stage schedule is reasonable: first train the speaker representation and decoder with a neutral style prior, then introduce reference style conditioning and adversarial separation. This order reduces the chance that style latents absorb speaker identity before the speaker classifier has become informative.

5.3 Inference-time control

At inference, the user supplies text, an optional set of attribute observations, and an optional style reference. The reliability gate computes e_spk from whatever attributes are available. If a style reference is supplied, the style encoder produces z_s; otherwise z_s is sampled from the prior. The decoder then generates acoustic features with adaptive normalization and gated cross-attention. The guidance scale can be adjusted to favor identity adherence, style adherence, or naturalness. Because the speaker and style pathways are separate, the system can generate the same text with a fixed voice and varying style, or with a fixed style and varying voice attributes.

A practical constraint is that short reference utterances provide weak estimates of pitch range and tempo. The gate should therefore down-weight attributes whose confidence estimates fall below a threshold, and the style encoder should rely more on the prior when the reference is short or has a low signal-to-noise ratio. This behavior is preferable to forcing a confident but unreliable embedding, which would produce unstable prosody and identity drift.

6. Analytical Assessment and Design Trade-offs

The proposed framework improves on monolithic speaker conditioning in interpretability and partial controllability. A single speaker vector is efficient and expressive, but it offers no direct way to adjust pitch range while preserving timbre or to request a slower tempo without changing vocal-tract characteristics. Attribute-specific embeddings expose these dimensions and allow the reliability gate to represent uncertainty. The cost is added model complexity and the need for attribute observations or reliable attribute classifiers. When attributes are unavailable, the gate reduces to a learned prior and the system behaves more like a conventional speaker-conditioned synthesizer.

Disentanglement remains the central technical risk. Orthogonality and adversarial losses reduce linear identity leakage, but they do not guarantee that style latents are semantically meaningful. A style latent may encode correlated channel properties, especially when the training corpus contains talkers recorded in distinct environments. Channel augmentation during training, such as random equalization and noise injection, can reduce this shortcut. However, aggressive augmentation can remove cues that are part of a talker’s natural voice, so augmentation strength should be treated as a design parameter rather than a fixed preprocessing step.

Adaptive style conditioning introduces a second trade-off between control strength and naturalness. Strong guidance improves adherence to the requested style but can produce overshoot in pitch and energy, especially for out-of-distribution references. Weak guidance preserves naturalness but may ignore the reference. Separate guidance scales for speaker and style are therefore preferable to a single scale, because identity and style have different reliability profiles. A style reference from a different talker may contain prosodic patterns that are incompatible with the target speaker's articulation; gated cross-attention mitigates this by admitting local style information only where it is consistent with the speaker embedding.

Evaluation of such a system should not rely on a single mean opinion score. Speaker similarity can be assessed with verification embeddings and equal error rate, style transfer with pitch and duration statistics and style classification accuracy, and naturalness with human listening tests. These metrics answer different questions and can conflict: a model that maximizes speaker similarity may suppress style variation, while a model that maximizes style transfer may drift in identity. The framework therefore implies a multi-objective evaluation protocol rather than a single aggregate score. The protocol should also report inference cost, because iterative diffusion decoding and cross-attention increase latency relative to feed-forward synthesis. Latent diffusion and reduced-step samplers can mitigate this cost, but they introduce approximation error that must be balanced against quality.

The framework has clear boundary conditions. It assumes that attribute observations or attribute classifiers are available during training and that reference style utterances are long enough to estimate prosodic statistics. It assumes that the training corpus contains sufficient speaker and style diversity for the gate and style encoder to learn separable representations. It also assumes that the decoder can absorb conditioning at multiple layers without destabilizing training. When these conditions fail, the system may still generate intelligible speech, but personalization and style control become less reliable. These limitations are not defects of the formulation; they define the operating regime in which multi-attribute conditioning is technically justified.

7. Conclusion

Personalized neural voice generation requires more than a high-quality speaker embedding and a style label. It requires a representation that exposes the attributes that users and applications actually want to control, and a conditioning mechanism that allows style to vary without rewriting identity. This manuscript developed a framework based on multi-attribute speaker embeddings and adaptive style conditioning. Attribute-specific encoders with a reliability gate support partial personalization and interpretable control; a style encoder with posterior regularization supports reference-based and prior-based style inference; adaptive normalization and gated cross-attention inject style and speaker information at different levels of the decoder; and a composite objective with adversarial and orthogonality terms reduces identity leakage. The analytical assessment shows that the main advantages are controllability, robustness to missing attributes, and support for style interpolation, while the main risks are attribute entanglement, channel bias, guidance instability, and added inference cost. These trade-offs define a technically coherent basis for personalized voice generation and a set of explicit conditions for evaluating whether multi-attribute conditioning improves over monolithic speaker representations.

References

1. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

2. 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 2017 (pp. 4006-4010). ISCA.

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

4. Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Advances in Neural Information Processing Systems 31. Curran Associates.

5. Wan, L., Wang, Q., Papir, A., & Lopez Moreno, I. (2018). Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4879-4883). IEEE.

6. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Proceedings of Interspeech 2020 (pp. 211-215). ISCA.

7. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., & 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. 4961-4970). PMLR.

8. Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., & Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In Proceedings of the 35th International Conference on Machine Learning (pp. 4693-4702). PMLR.

9. Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (pp. 5530-5540). PMLR.

10. Zhang, Y., Jiang, Z., Li, R., Pan, C., He, J., Huang, R., ... & Zhao, Z. (2024, November). Tcsinger: Zero-shot singing voice synthesis with style transfer and multi-level style control. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 1960-1975).

11. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021). Grad-TTS: A diffusion probabilistic model for text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (pp. 8823-8833). PMLR.

12. Kong, J., Ping, W., Huang, L., Zhao, K., & Catanzaro, B. (2021). DiffWave: A versatile diffusion model for audio synthesis. In International Conference on Learning Representations.

13. Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Proceedings of the 39th International Conference on Machine Learning (pp. 3015-3028). PMLR.

14. Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., & Hsu, W.-N. (2023). Voicebox: Text-guided multilingual universal speech generation at scale. In Advances in Neural Information Processing Systems 36. Curran Associates.

15. Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 5288-5295.

16. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems 33 (pp. 6840-6851). Curran Associates.

17. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In Advances in Neural Information Processing Systems 30. Curran Associates.

18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684-10695). IEEE.