Controllable Audio Generation for Virtual Characters Based on Speaker Identity and Expressive Speech Modeling

Oonagh A. Brien
Department of Computer Science, University of Houston, Houston, TX, USA.
oonaghbrien695@uh.edu

Gianna Borghese
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
giannawork@ku.edu

Abstract

Virtual characters require voices that remain recognizably consistent across utterances while allowing deliberate control over emotion, speaking rate, emphasis, and other expressive qualities. Existing neural speech generators can produce natural audio, yet identity preservation and expressive controllability are often entangled: changing prosody can alter perceived speaker timbre, while enforcing a target identity can suppress stylistic variation. This article develops a methodological framework for controllable audio generation that separates a speaker identity embedding from a low-dimensional expressive speech representation and conditions a neural acoustic generator on both factors plus interpretable control variables. The framework combines discriminative speaker modeling, self-supervised speech representations, style-based conditioning, variational regularization, adversarial naturalness constraints, and covariance-based disentanglement. It also specifies a two-stage inference procedure in which candidate generations are filtered by speaker similarity, expressive agreement, and uncertainty. The analytical contribution is a precise account of how identity invariance, expressive range, and controllability trade off under limited labeled data and nonparallel references. The discussion identifies failure modes including identity leakage through prosody, expression-induced timbre drift, control entanglement, and evaluation mismatch. The resulting framework supports virtual character pipelines that demand stable identity across content and controlled variation within character-specific expressive bounds, and it clarifies which design choices require empirical calibration rather than architectural novelty alone.

Keywords

controllable audio generation; speaker identity modeling; expressive speech synthesis; virtual characters; disentangled representation; prosody control

1. Introduction

Virtual characters in games, interactive narrative systems, and embodied conversational agents need more than intelligible speech. They need a stable vocal identity that listeners recognize across sentences, emotional states, and interaction contexts, together with deliberate control over expressive behavior. A character may be required to sound warm but hurried in one scene, restrained and ironic in another, and authoritative without losing the same

underlying timbre. This requirement turns speech generation into a constrained modeling problem: the generated audio must vary along interpretable expressive dimensions while remaining anchored to a speaker identity that is not overwritten by those variations. Neural speech synthesis has advanced enough that naturalness alone is no longer the central obstacle [1], [2], [3]. The persistent difficulty is controlled variation under identity consistency.

Statistical parametric synthesis established the value of explicit acoustic modeling and controllable vocoding, but its intermediate representations limited expressive nuance [1]. Neural waveform models then improved the fidelity of generated audio [2], and end-to-end text-to-speech systems reduced reliance on hand-engineered pipelines [3]. In parallel, speaker verification research produced discriminative embeddings that can represent identity for enrollment and retrieval [4]. These developments suggest a natural architecture for virtual characters: use an identity encoder to represent who is speaking, an expressive encoder to represent how the utterance is delivered, and a generator that combines both with user-facing controls. The difficulty is that identity and expression are not statistically independent in natural speech. Emotional arousal changes pitch range, energy, and spectral tilt; speaking rate changes rhythmic structure; emphasis changes local duration and intensity. A model trained only for reconstruction can preserve these correlations, but it may also learn to use prosodic cues as shortcuts for identity, which undermines control.

A focused design gap follows. Prior work has demonstrated style transfer, prosody transfer, and unsupervised style tokens, but these approaches do not by themselves specify how a virtual character's identity should remain invariant under expressive editing or how controllability should be assessed when labels are sparse and references are nonparallel [5], [6], [7]. Style transfer for out-of-domain singing voice synthesis, for example, addresses a related problem of preserving target timbre while applying stylistic variation [5]. Prosody transfer systems show that reference prosody can be transferred across speakers, but they also expose leakage between speaker characteristics and prosodic style [6]. Unsupervised style tokens provide global control over speaking style, yet the tokens are not guaranteed to align with interpretable attributes such as emotion intensity, rate, or emphasis [7]. For virtual characters, the relevant question is not merely whether a model can imitate a reference. It is whether a character's voice can be edited along defined dimensions while remaining the same character.

This article develops a methodological framework for controllable audio generation that treats speaker identity and expressive speech as separate but interacting factors. The central contribution is a conditional generation formulation in which a speaker embedding, a low-dimensional expressive representation, and an interpretable control vector jointly determine the generated acoustic sequence. The framework combines a discriminative identity objective, self-supervised speech representations, style-based conditioning, variational regularization, adversarial naturalness constraints, and a covariance-based disentanglement term. It further specifies a two-stage inference procedure that filters candidate outputs by identity similarity, expressive agreement, and uncertainty. The analysis does not report a completed benchmark study; instead, it derives design consequences, boundary conditions, and failure modes from the structure of the problem and from established results in speech synthesis, speaker verification, and generative modeling. This framing is appropriate because the core research question concerns how to formulate and constrain controllable generation for virtual characters, not whether a single architecture improves a leaderboard score.

2. Related Work and Problem Formulation

The technical lineage begins with parametric synthesis, where acoustic features such as spectral envelopes, fundamental frequency, and duration are predicted from text and then rendered by a vocoder [1]. That paradigm made control explicit, but it also separated naturalness from controllability in ways that limited expressive range. Neural autoregressive waveform generation changed the fidelity ceiling by modeling raw audio distributions directly [2]. End-to-end systems then learned alignments and acoustic representations from text-audio pairs, reducing the need for brittle feature engineering [3]. These advances provide the generative capacity for virtual character audio, but they do not resolve the identity-expression entanglement problem.

Speaker identity modeling supplies one necessary component. Discriminative embeddings trained with verification objectives can separate speakers in an embedding space, and generalized end-to-end losses make such embeddings useful for enrollment from few utterances [4]. For character voices, an identity embedding must be invariant to phonetic content and expressive state. That invariance is not automatic. An embedding trained only to classify speakers may encode recording conditions, habitual pitch range, or emotional tendencies as identity cues. The problem is analogous to style transfer in singing voice synthesis, where a target timbre must be preserved under a new stylistic interpretation [5]. The difference for virtual characters is that expressive control is not only a transfer target; it is an interactive parameter that must be stable across many generated utterances.

Prosody transfer and style modeling provide a second line of work. End-to-end prosody transfer demonstrates that reference prosody can be separated from text and transferred to a target speaker, yet residual speaker information can leak through the prosodic representation [6]. Unsupervised style tokens offer global control over style without explicit labels, but their interpretability is limited and their axes are learned from data rather than defined by an application [7]. Style-based generative architectures show that affine modulation of intermediate features can provide strong global control over generation [8]. That idea is relevant to virtual characters because it permits expressive attributes to modulate a shared identity-conditioned generator without replacing the identity pathway. The open question is how to constrain the modulation so that identity remains stable and controls remain semantically meaningful.

Self-supervised speech representations add a complementary source of structure. HuBERT learns discrete and continuous representations through masked prediction of hidden units [9], and wav2vec 2.0 learns representations by contrastive prediction over quantized latent speech representations [10]. Such features capture phonetic and prosodic information without requiring transcript labels for every utterance. They are attractive for expressive modeling because they can represent fine-grained temporal variation, but they are not inherently disentangled. A representation that predicts masked units may encode speaker, channel, and affect simultaneously. For controllable virtual characters, self-supervised features are best treated as a rich input to a constrained expressive encoder rather than as a ready-made control space.

Modern end-to-end generators provide the synthesis backbone. Variational autoencoder with adversarial learning for text-to-speech combines a conditional prior, a posterior encoder, and a adversarial waveform decoder [11]. Non-autoregressive models such as FastSpeech 2 improve inference efficiency and expose duration, pitch, and energy as explicit predictors [12]. Zero-shot multi-speaker systems show that speaker embeddings can condition a shared

generator and support voice conversion with limited target data [13]. These systems establish that conditioning on a speaker representation is feasible, but they typically optimize naturalness and speaker similarity rather than interpretable expressive editing. A virtual character pipeline needs both: the generator must render a stable identity and expose controls that a director, game logic, or dialogue system can set.

The problem can be stated more precisely. Let x be a reference utterance from a virtual character, c a control vector whose components specify expressive intent such as emotion category, intensity, rate, and emphasis, and y a phonetic or textual content sequence. The goal is to generate audio $\hat{x}$ that matches the content y, preserves the identity of x, and realizes the expressive intent c. The principal challenge is that identity and expression are correlated in the training distribution. If the model minimizes reconstruction error alone, it can achieve low error by copying expressive cues from the reference even when c requests a different delivery. If it over-constrains identity, it can remove expressive variation that makes a character feel alive. The framework below addresses this tension by separating the reference-dependent expressive representation from the control vector and by adding explicit disentanglement and inference-time selection.

3. Methodological Framework

3.1 Speaker Identity Representation

The identity pathway begins with an encoder that maps a reference utterance to a speaker embedding. The reference may be a short enrollment sample, a previously generated line, or a stored character voice profile. The encoder should be trained with a speaker classification or verification objective so that embeddings from the same character cluster together and embeddings from different characters separate. A discriminative formulation is:

$$e_s = f_s(x; \theta_s)$$

Here e_s is the speaker embedding, f_s is the identity encoder with parameters θ_s, and x is the reference waveform or its mel-spectrogram. The embedding is used as a conditioning variable for the generator, not as a waveform template. This distinction matters because a template-based approach would preserve identity by copying acoustic detail, but it would also copy the expressive state of the reference. The encoder should therefore be trained on utterances that vary in content and expression so that the embedding is forced to represent stable vocal characteristics.

A classification loss can encourage separability across characters:

$$Lspk = -\sum_k y_k \, log \, softmax(W \, e_s)_k$$

In this expression, k indexes speaker classes, y_k is a one-hot speaker label for the reference utterance, W is a linear classification matrix, and $\text{softmax}(W\, e_s)_k$ is the predicted probability of class k. The loss is not sufficient by itself. It must be paired with augmentation or sampling that varies content and expression within each character; otherwise the encoder can learn expressive correlates as identity features. Self-supervised representations from models such as HuBERT and wav2vec 2.0 can supply robust frame-level features to the encoder [9], [10], but the final embedding should be pooled and regularized so that short-term prosodic fluctuations do not dominate the identity code.

3.2 Expressive Speech Modeling

Expressive speech is represented by a low-dimensional latent variable z_e that captures residual delivery characteristics not explained by content and identity. The generator receives the identity embedding e_s, the expressive latent z_e, and the control vector c. The control vector contains application-facing attributes such as emotion category, intensity, speaking rate, emphasis strength, and pause style. The generator produces an acoustic representation or waveform:

$$\hat{x} = G(e_s, f_e(x; \theta_e), c)$$

Here G is the conditional generator, f_e is the expressive encoder with parameters θ_e, and $\hat{x}$ is the generated audio. The expressive encoder is not asked to produce the final control vector. It produces a latent that captures fine-grained correlates of delivery, while c provides the interpretable direction. This separation allows the system to use a reference utterance as a starting point without treating the reference as an unchangeable style label.

The control vector should be structured rather than a single global token. Emotion category can be represented as a categorical variable, intensity as a scalar, rate as a multiplicative or additive duration factor, and emphasis as a sequence-level annotation or phoneme-level weight. The generator can inject these variables through affine modulation of intermediate features, following the principle that global style can modulate a shared generative pathway [8]. For a virtual character, this design has a practical advantage: the same identity embedding can be reused across all lines, while the control vector is set per utterance by dialogue logic. A character can therefore remain recognizable when the game requests anger, fatigue, or urgency without requiring a separate model for each emotional state.

Expressive modeling also needs a representation of prosodic structure. Prosody is not a single scalar; it includes pitch accents, boundary tones, rhythm, and voice quality, and these dimensions interact with syntax and pragmatics [15], [16]. A model that controls only mean pitch and speaking rate will produce audible but shallow variation. A more adequate expressive encoder should capture local pitch contours, energy envelopes, and duration patterns, while the control vector specifies the intended direction at a coarser level. The latent z_e can be regularized toward a prior conditioned on c so that the generator learns to associate control settings with plausible expressive trajectories. Emotional speech corpora such as IEMOCAP provide evidence that emotion categories are expressed through correlated changes in pitch, energy, and timing rather than isolated acoustic features [14]. That correlation is precisely why disentanglement constraints are necessary.

3.3 Controllable Generation Architecture

The generator can be organized as a shared identity-conditioned backbone with several conditioning interfaces. The identity embedding enters early and is repeated or projected into multiple layers so that timbre is preserved throughout generation. The control vector enters through affine modulation or feature-wise transformation at selected layers so that expressive settings influence local dynamics without overwriting the identity pathway. The expressive latent enters as a stochastic variable that captures residual variation not specified by c. A non-autoregressive acoustic model with explicit duration, pitch, and energy predictors is attractive for virtual characters because it supports predictable inference cost and direct manipulation of timing [12]. A conditional variational autoencoder with adversarial waveform generation is also compatible with this structure and can produce high-fidelity audio from compact latent representations [11]. Zero-shot multi-speaker conditioning suggests that a shared generator can generalize to unseen speakers when the identity embedding is informative [13].

The architecture should not collapse the expressive latent and the control vector into a single pathway. If c is the only expressive input, the model cannot represent within-category variation; all angry lines would be identical in delivery. If z_e is the only expressive input, the application cannot set interpretable controls. The dual pathway lets c define the intended region of expressive space and z_e supply natural variation within that region. During training, z_e is inferred from the reference utterance. During inference, z_e can be sampled from the conditional prior, taken from a reference performance, or interpolated between the two. This flexibility is useful for virtual characters because a director may want to preserve the timing of a recorded line while changing its emotional color, or preserve the emotional color while changing timing.

3.4 Training Objectives and Disentanglement Constraints

Training must balance reconstruction, identity preservation, expressive control, naturalness, and disentanglement. A composite objective can be written as:

$$Lgen = Lrec + \lambda_1 Lspk + \lambda_2 Ladv + \lambda_3 DKL + \lambda_4 Ldis$$

Lrec is a reconstruction loss between the generated and target acoustic features or waveforms. Lspk is the speaker classification or verification loss applied to embeddings extracted from generated audio. Ladv is an adversarial naturalness loss. DKL is a Kullback-Leibler divergence between the posterior expressive distribution and a prior conditioned on c. Ldis is a disentanglement penalty. The coefficients λ_1 through λ_4 control the trade-off among these terms. The variational term follows the standard autoencoding variational Bayes formulation [17], while the adversarial term follows generative adversarial training with stabilized discriminator objectives [18]. The speaker term should be applied to generated audio through a frozen or slowly updated identity encoder so that the generator cannot satisfy the loss by producing embeddings that are discriminative but acoustically implausible.

A simple covariance penalty can discourage linear dependence between identity and expressive factors within a batch:

$$Ldis = \|Cov(e_s, z_e)\|_1$$

Cov denotes the covariance matrix between speaker embeddings and expressive latents computed over the batch, and the entrywise L1 norm penalizes nonzero cross-covariance. This penalty is not a complete solution to nonlinear dependence, but it is analytically transparent and computationally cheap. It should be supplemented by adversarial disentanglement or mutual-information estimation when the expressive latent is high-dimensional. Variance-weighted regularization in the spirit of beta-VAE can further encourage axis alignment and informativeness in z_e [20]. The important design principle is that disentanglement is enforced at the level of the factors the application actually uses. Penalizing covariance between e_s and z_e is meaningful only if z_e is trained to predict expressive attributes and e_s is trained to be stable across them.

The adversarial naturalness term also requires care. A discriminator that sees only real and generated waveforms can reward identity-consistent but expressively flat outputs if the training data contain many neutral utterances. Conditioning the discriminator on the control vector c encourages it to assess naturalness within the requested expressive region. Improved training techniques for generative adversarial networks reduce instability, but they do not remove the need to balance adversarial and reconstruction losses [18]. Classifier-free guidance can be used during inference to sharpen adherence to c, at the cost of reduced

diversity [19]. For virtual characters, that trade-off is often acceptable when a dialogue system requests a specific emotion, but it is undesirable when the goal is to generate varied background chatter.

4. Analytical Assessment and Design Trade-offs

4.1 Controllability and Identity Preservation

The central trade-off is between identity invariance and expressive range. If the speaker embedding is too dominant, the generator may ignore c and reproduce the reference delivery. If the control pathway is too dominant, the generator may realize the requested expression but drift away from the character's timbre. A useful way to reason about this trade-off is to separate interpolation from extrapolation. Interpolation within the range of expressive variation observed for a character is likely to preserve identity because the generator has seen similar acoustic configurations. Extrapolation to an emotion intensity or speaking rate far outside the training range is riskier because the acoustic correlates of extreme expression can overlap with identity cues. A whispered line, for example, removes much of the periodic voicing that listeners use to identify a speaker. No conditioning scheme can fully recover identity information that is absent from the target acoustic style, so the system should expose a controllability envelope within which identity is expected to remain stable.

Control editing can be formulated as a displacement in expressive latent space:

$$\hat{x}' = G(e_s, z_e + \alpha \Delta c, c')$$

Here α is a scalar step size, Δc is a direction associated with a control change, and c′ is the updated control vector. The equation makes explicit that control is not a binary switch. Small α values produce subtle variation, while large values move farther from the reference delivery. The direction Δc should be estimated from labeled contrasts or from a supervised control regressor, not chosen arbitrarily. If the same Δc is used for all characters, the system assumes that expressive directions are shared across speakers. That assumption is convenient but often false: a character with a low habitual pitch may realize anger through increased vocal effort rather than a large upward pitch shift. The framework should therefore allow character-specific calibration of Δc while keeping the control semantics global.

4.2 Data Efficiency and Generalization

Virtual character projects rarely have large labeled corpora for every character and every expressive state. The framework must therefore exploit nonparallel data. The identity encoder can be trained on many speakers with speaker labels, the expressive encoder can be trained on emotional corpora with weak or categorical labels, and the generator can be trained on multi-speaker data with content and speaker conditioning. This division reduces the burden on any single dataset. It also introduces a domain mismatch: a character voice may be a studio recording with a specific microphone and acting style, while the expressive corpus may contain acted emotional speech from a different recording environment. The model should not treat recording channel as identity or as expressive style. Channel augmentation and adversarial domain classifiers can reduce that risk, but they add hyperparameters and can remove cues that are genuinely part of a character's vocal identity.

Generalization to out-of-domain expression is another boundary condition. A model trained on discrete emotion categories may not represent blended or subtle states such as "politely impatient" or "tired but amused." Continuous control variables and latent expressive references are better suited to such cases than a small set of categorical labels. The cost is

interpretability: a continuous control space is harder for a director to set without a calibration interface. A practical compromise is to expose a small number of named controls while allowing a reference-performance latent to supply residual nuance. That design preserves the character identity through e_s and gives the application both coarse and fine expressive handles.

4.3 Evaluation Logic and Error Modes

Evaluation for this problem cannot rely on naturalness alone. A generated line may sound natural yet fail to preserve the character's identity or fail to realize the requested emotion. A defensible protocol should separate at least four dimensions: intelligibility, identity similarity, expressive agreement, and controllability. Intelligibility can be assessed with automatic speech recognition error and human transcription. Identity similarity can use a speaker verification embedding and listener judgments, but automatic scores should not be treated as ground truth because expressive extremes can reduce verification accuracy even when listeners recognize the character. Expressive agreement can use a combination of human ratings and a trained attribute classifier, with the caveat that classifiers inherit the label biases of their training data. Controllability requires paired comparisons: for a fixed content and identity, changing c should produce a measurable change in the intended expressive attribute and a smaller change in unrelated attributes.

Candidate selection can formalize these considerations:

$$\hat{x} * = argmax_k \left[\lambda_i \, s(G_k, e_s) \; - \; \lambda_p \, d(G_k, c) \; - \; \lambda_n \, u(G_k)\right]$$

G_k is the kth candidate generation, s is a speaker similarity score, d is an expressive deviation between the candidate and the requested control, u is an uncertainty estimate, and λ_i, λ_p, and λ_n are weights. The uncertainty term can be estimated from variance across multiple stochastic samples or from predictive entropy. This selection rule is not a substitute for training-time disentanglement, but it provides a safety layer for interactive systems. It can reject candidates that drift in identity or undershoot the requested expression before they reach the virtual character's dialogue loop.

Several error modes follow directly from the framework. Identity leakage through prosody occurs when the expressive latent encodes speaker-specific pitch range or vocal effort; the covariance penalty and adversarial disentanglement reduce but do not eliminate this. Expression-induced timbre drift occurs when strong controls push the generator outside the character's vocal manifold. Control entanglement occurs when changing rate also changes pitch or emotion, which can happen if the control variables are correlated in training data. Evaluation mismatch occurs when automatic identity or emotion classifiers reward different cues than human listeners. These are not peripheral implementation issues; they define the practical limits of controllable generation for virtual characters.

5. Technical Discussion and Implications

The framework's main implication is that virtual character audio generation should be treated as a constrained factor modeling problem, not only as a text-to-speech problem with a speaker embedding appended. The identity embedding must be trained for invariance, the expressive latent must be trained for controlled variation, and the generator must be trained to respect both under nonparallel data. This division of labor has architectural consequences. It favors a shared generator with multiple conditioning interfaces over separate per-emotion models, because a shared generator can reuse identity information and generalize to unseen combinations. It also favors explicit control variables over purely unsupervised style tokens

for application-facing control, because a game engine or narrative system cannot reliably set a token index without a semantic interface. Unsupervised latent variation can still be useful for within-category nuance, but it should not be the only expressive pathway.

The second implication concerns evaluation and deployment. A virtual character pipeline needs an online monitor for identity and expressiveness, not just offline naturalness scores. The candidate selection rule above is a minimal version of such a monitor. In a real system, the monitor would maintain a character-specific identity reference, track drift across a dialogue, and flag requests that fall outside the controllability envelope. This is especially important for long-form interaction, where small identity drifts can accumulate. A line that sounds acceptable in isolation may sound like a different character after several exchanges. The framework therefore benefits from character-level enrollment and periodic re-anchoring rather than relying on a single global embedding.

The third implication is that controllability is inherently a data and calibration problem. No loss function can guarantee that a learned control axis corresponds to a director's intended emotion if the training labels do not distinguish that emotion from correlated acoustic changes. The most useful technical work may therefore lie in constructing contrastive training pairs that isolate one expressive factor at a time, and in designing calibration procedures that map application controls to character-specific latent displacements. This does not diminish the role of generative architecture. It clarifies that architecture creates the capacity for control, while data design and calibration determine whether that capacity is usable.

6. Conclusion

This article presented a methodological framework for controllable audio generation for virtual characters that separates speaker identity from expressive speech and conditions generation on interpretable control variables. The framework combines a discriminative identity encoder, a self-supervised or spectral expressive encoder, a style-conditioned generator, variational and adversarial training objectives, and a covariance-based disentanglement penalty. It also specifies a control-editing formulation and a candidate selection rule that accounts for identity similarity, expressive deviation, and uncertainty. The analysis shows that the central difficulty is not generating natural audio in isolation but maintaining a character's identity while allowing controlled expressive variation. Identity invariance, expressive range, and control fidelity trade off in ways that depend on the training distribution, the controllability envelope, and the degree to which expressive cues overlap with identity cues. Disentanglement constraints reduce leakage but cannot remove the need for character-specific calibration and evaluation. For virtual character pipelines, the practical consequence is that controllable generation should be designed as a factorized system with explicit controls, a stable identity anchor, and an online monitor for drift. This framing provides a technically grounded basis for future empirical work without assuming that naturalness alone is a sufficient objective.

References

1. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039–1064.

2. 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.

3. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Ying, Z., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.

4. 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.

5. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

6. 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. arXiv preprint arXiv:1803.09047.

7. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, Y., 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 (PMLR 80, pp. 5180–5189).

8. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401–4410).

9. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460.

10. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449–12460).

11. 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 (PMLR 139, pp. 5530–5540).

12. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. 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 (PMLR 162, pp. 2709–2720).

14. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359.

15. Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In J. Tao & T. Tan (Eds.), Affective information processing (pp. 111–126). Springer.

16. Wagner, P., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. Language and Cognitive Processes, 25(7–9), 905–945.

17. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In International Conference on Learning Representations.

18. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In Advances in Neural Information Processing Systems (Vol. 29).

19. Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.

20. Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations.