Reference-Guided Acoustic Representation Learning for Personalized Music and Speech Generation

Logan D. Blanchard
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
logandblanchard@unr.edu

Teodoro Palma
School of Computing, Clemson University, Clemson, SC, USA.
teodoro.work@clemson.edu

Natasza Bulawa
Department of Computer Science, George Mason University, Fairfax, VA, USA.
bulawa1971@gmu.edu

Abstract

Reference-guided acoustic generation seeks music or speech that follows the timbre, prosody, and expressive style of a short audio prompt while remaining controlled by text, lyrics, phonemes, or a musical score. A central difficulty is that reference audio simultaneously carries stable personalization cues, transient performance choices, content residue, and recording conditions; models that compress the reference into a single vector often entangle these factors and transfer them indiscriminately. This paper develops a representation-learning framework that factorizes reference audio into style, prosody, timbre, and speaker or instrument subspaces, links these subspaces to symbolic content through cross-attention and feature-wise modulation, and controls leakage with information-bottleneck and mutual-information penalties. The framework formalizes reference-conditioned generation as a marginal likelihood over latent reference factors, specifies a training objective that combines reconstruction, adversarial realism, content alignment, disentanglement, and posterior regularization, and introduces a perturbation-smoothed likelihood for inference when reference quality or domain match is uncertain. Analytical assessment identifies trade-offs among fidelity, personalization strength, and content leakage, clarifies why unsupervised disentanglement alone is insufficient without weak inductive biases, and distinguishes failure modes in speech and singing or instrumental music. The contribution is a coherent design account and evaluation logic for personalized acoustic generation under reference uncertainty, with implications for controllable synthesis, model calibration, and responsible deployment.

Keywords

reference-guided generation; acoustic representation learning; personalized speech synthesis; singing voice synthesis; disentanglement; uncertainty-aware inference

1. Introduction

Reference-guided acoustic generation differs from conventional text-to-speech or symbolic music generation because the target output is conditioned jointly on symbolic content and on an audio example whose relevant attributes are not explicitly annotated. Neural generative models provide the basic machinery for learning such conditional

distributions [1], [2], while attention over symbolic sequences and autoregressive waveform modeling supply the two dominant architectural ingredients for high-fidelity synthesis [3], [4]. In speech, end-to-end systems demonstrated that text can be mapped directly to acoustic features or waveforms [5], [6], and non-autoregressive designs later improved controllability and inference efficiency [7]. Expressive prosody transfer and unsupervised style tokens showed that a reference utterance can shape rhythm, emphasis, and global speaking style beyond speaker identity [8], [9]. Surveys of neural speech synthesis document the breadth of these developments but also expose the lack of a unified account of how reference attributes should be represented, separated, and evaluated [10].

Singing voice synthesis makes the problem sharper because pitch, timing, timbre, and expressive style interact continuously. Style transfer for out-of-domain singing voice synthesis demonstrates that reference style can be transferred beyond a narrow training distribution, yet it also illustrates the risk of entangling singer identity with pronunciation, recording condition, and musical interpretation [11]. Music generation systems based on symbolic transformers, multi-track adversarial training, and large-scale audio modeling have expanded the range of controllable musical material [12], [13], [14], and neural audio autoencoders have shown that waveform-level synthesis can be learned from musical data [15]. These lines of work motivate a representation-level treatment of personalization: a reference should be transformed into interpretable acoustic factors that can be selected, composed, and regularized rather than copied as an undifferentiated embedding.

The knowledge gap addressed here concerns the relationship between reference representation, conditioning, and leakage control. Existing systems often rely on a global reference embedding, sometimes augmented with style tokens or prosody encoders, and then optimize reconstruction and adversarial realism. Such designs can produce convincing outputs when reference and target are similar, but they do not specify which reference factors should persist under domain shift, how content residue enters the conditioning path, or how uncertainty in the reference posterior should influence decoding. Representation learning for audio offers partial remedies. Information-bottleneck decomposition can separate speech factors [16], discrete latent models provide compact codes [17], neural audio compression supplies high-fidelity analysis-synthesis components [18], and self-supervised speech encoders provide transferable acoustic structure [19], [20]. Calibration research further shows that modern neural predictors can be overconfident [21], and disentanglement studies caution that unsupervised separation is not identifiable without inductive bias or supervision [22]. These findings imply that personalized acoustic generation needs an explicit factorization and a leakage-aware training logic, not only a larger reference encoder.

The central contribution of this paper is a reference-guided acoustic representation learning framework for personalized music and speech generation. The framework factorizes a reference signal into style, prosody, timbre, and speaker or instrument subspaces; couples the selected factors to symbolic content through cross-attention and feature-wise modulation; and trains the model with reconstruction, adversarial realism, content alignment, mutual-information, and posterior regularization terms. It also introduces a perturbation-smoothed likelihood to account for reference uncertainty and provides an analytical assessment of fidelity, personalization, leakage, and domain-specific failure modes. The argument is methodological and analytical: it specifies design choices, derivable consequences, and conditions under which the proposed representation is appropriate, without reporting synthetic experiments or empirical scores.

2. Related Work and Problem Setting

Acoustic representation learning has moved from hand-designed spectral features to learned latent spaces. Information-bottleneck methods can encourage separation of speech content from speaker or channel factors [16], while vector-quantized autoencoders produce discrete acoustic codes suitable for sequence modeling [17]. Neural audio compression has advanced high-fidelity waveform reconstruction at reduced bit rates [18], and self-supervised encoders such as wav2vec 2.0 and HuBERT learn representations that transfer across speech tasks [19], [20]. These developments suggest that a reference encoder should not be trained only as a style classifier; it should preserve enough acoustic detail for synthesis while exposing factors that can be regularized against content leakage.

Expressive synthesis has supplied several conditioning mechanisms. Prosody transfer conditions an autoregressive model on reference prosody [8], style tokens provide an unsupervised global style representation [9], and surveys of neural speech synthesis summarize the shift toward controllable and end-to-end systems [10]. Singing voice synthesis adds score-dependent pitch and timing, and style transfer for out-of-domain singing shows that expressive reference information can be useful beyond a single domain [11]. Music generation work has explored long-range symbolic structure [12], multi-track symbolic generation [13], large-scale audio generation [14], and neural waveform synthesis for musical notes [15]. Together, these systems establish that reference conditioning is feasible, but they do not fully specify how to prevent a reference from transferring content, noise, or recording artifacts into a target with different symbolic content.

Two further strands inform the proposed design. First, alignment methods such as connectionist temporal classification provide a way to preserve symbolic content without frame-level labels [23], and attention-based speech recognition shows how soft alignment can connect variable-length sequences [25]. Second, adversarial objectives, especially Wasserstein formulations, offer a way to encourage realistic outputs when likelihood-based reconstruction tends to oversmooth [24]. Calibration work is relevant because a reference encoder that produces a sharp posterior may ignore plausible alternative interpretations of a noisy prompt [21]. Disentanglement theory is also relevant because it warns that unsupervised separation of factors is generally non-identifiable [22]; the present framework therefore treats factorization as an inductive bias supported by weak supervision, content-adversarial terms, and architectural constraints rather than as a guaranteed consequence of training.

The problem setting can be stated compactly. Let x denote symbolic content, such as a phoneme sequence, lyrics plus pitch contour, or a musical score. Let R denote a reference audio segment, and let y denote the acoustic target, represented as a waveform or a spectral sequence. The model should generate y that realizes x while matching selected personalization attributes of R, such as voice quality, instrument timbre, expressive timing, or singing style. The model should not copy the reference content, background noise, or channel characteristics unless those attributes are explicitly requested. This distinction between requested personalization and nuisance transfer is the central design constraint.

3. Reference-Guided Acoustic Representation Learning Framework

3.1 Generative Formulation and Reference Factorization

The reference-conditioned generation problem is formulated as a marginal likelihood over latent reference factors. The target acoustic sequence y is generated from symbolic content x

and reference audio R by integrating over a latent variable z that summarizes the reference. The likelihood is

$$p\theta(y \mid x, R) = \int p\varphi(z \mid R)\, p\theta(y \mid x, z)\, dz$$

where θ denotes generation parameters, φ denotes reference-encoder parameters, pφ(z | R) is the reference posterior or encoder distribution, and pθ(y | x, z) is the content-conditioned acoustic generator. This expression makes the role of uncertainty explicit: when R is short, noisy, or domain-mismatched, pφ(z | R) should remain broad over plausible personalization factors rather than collapsing to a single point estimate.

The reference encoder Eφ maps R into four subspaces, concatenated as

$$E\varphi(R) = [z_s; z_p; z_t; z_i]$$

where z_s denotes style and expressive performance, z_p denotes prosody or micro-timing, z_t denotes timbre and spectral envelope, and z_i denotes speaker or instrument identity. The separation is conceptual and architectural rather than absolute: timbre and identity overlap, and prosody and style overlap, especially in singing. The purpose of the factorization is to give each factor a distinct conditioning route and a distinct regularization target, not to assert that the underlying physical causes are statistically independent.

The generator receives x and the selected factors. For speech, z_i might control speaker identity while z_p controls rhythm and emphasis; for music, z_i might control instrument identity while z_t controls timbre and z_s controls articulation and expressive timing. A domain indicator can select which factor subset is active, but the framework avoids hard-coded disentanglement by allowing soft gates over the four subspaces. This design acknowledges that a reference may be informative about prosody even when speaker identity is irrelevant, and vice versa.

3.2 Disentanglement and Content Leakage Control

The main threat to personalization is leakage: a reference encoder can encode phonetic content, score content, lyrics, background noise, or channel effects into a factor intended for style or timbre. If such a factor conditions the generator, the output may reproduce reference content or artifacts. The framework controls leakage through mutual-information penalties and content-adversarial regularization. A general penalty is

$$L_M I = \sum_{a \neq b} I(z_a; z_b) + \sum_a I(z_a; c_R)$$

where a and b index the four latent subspaces, I(· ; ·) denotes mutual information, and c_R denotes a content residual extracted from the reference by a separate content encoder. The first term discourages redundancy among style, prosody, timbre, and identity; the second discourages each factor from carrying reference content. In practice, mutual information is estimated with variational bounds or contrastive estimators, and the penalty is weighted against reconstruction so that excessive separation does not destroy synthesis quality.

This objective reflects a deliberate theoretical stance. Disentanglement is not assumed to arise from an autoencoder alone, because unsupervised factor separation is generally non-identifiable without inductive bias [22]. The penalty therefore acts as a soft constraint, while architectural choices such as separate encoders, temporal pooling for global style, and frame-level processing for prosody provide additional structure. Weak supervision, such as speaker labels when available or score alignment for singing, can further anchor the subspaces. When

such supervision is unavailable, the model should be interpreted as producing partially separated factors with residual correlations, not as recovering true causal variables.

3.3 Cross-Attentive Guidance and Personalization Conditioning

Once reference factors are extracted, the generator must apply them to symbolic content without flattening temporal structure. Global style tokens can be broadcast across time, but prosody and timing require temporal alignment between reference and target. The framework uses cross-attention from target content positions to reference frames, followed by feature-wise modulation. A compact form of the conditioning update is

$$h_t = \gamma(z_ref) \odot f\theta(x_t, c_t) + \beta(z_ref)$$

where h_t is the hidden state at target position t, x_t is the local symbolic content, c_t is contextual content information, fθ is a content encoder, γ and β are modulation parameters produced from the selected reference factors z_ref, and ⊙ denotes elementwise multiplication. The factor z_ref may combine z_s, z_p, z_t, and z_i according to the synthesis task. Feature-wise modulation gives global factors such as timbre and speaker identity a stable influence, while cross-attention supplies position-dependent guidance for prosody and expressive timing.

For music, the content encoder may operate on note events, durations, and pitch; for speech, it may operate on phonemes and predicted durations. The cross-attention mechanism must be constrained so that it cannot copy reference content. One option is to detach attention weights from acoustic detail associated with content and to train a content-adversarial classifier on the attended representation. Another is to restrict attention to factor-specific reference frames selected by a content-invariance score. These constraints are not cosmetic: without them, attention can become a copy channel, especially when reference and target share lexical or melodic material.

3.4 Training Objectives and Optimization Logic

The model is trained with a composite objective that balances acoustic reconstruction, perceptual realism, content preservation, disentanglement, and posterior regularization. A representative form is

$$L = L_rec + \lambda_a dv L_a dv + \lambda_c tc L_c tc + \lambda_M I L_M I + \lambda_K L L_K L$$

where L_rec is a spectral or waveform reconstruction loss, L_adv is an adversarial realism term, L_ctc is a content-alignment loss, L_MI is the leakage penalty above, and L_KL regularizes the reference posterior toward a prior or toward a factor-specific prior. The coefficients λ_adv, λ_ctc, λ_MI, and λ_KL control the trade-offs. The adversarial term can use a Wasserstein objective to provide smoother gradients than a saturating minimax loss [24]. The content term can use connectionist temporal classification when frame-level alignments are unavailable [23], or a score-alignment loss when symbolic timing is known.

The coefficients are not free tuning knobs without consequences. Increasing λ_MI reduces leakage and may weaken expressive fidelity if style and timbre are naturally correlated. Increasing λ_adv improves realism but can amplify artifacts when the reference posterior is broad. Increasing λ_ctc strengthens content preservation but may suppress prosodic variation if the content encoder absorbs style information. The objective therefore encodes a design position: personalization should be strong only for the factors that are both identifiable and requested by the task. This position distinguishes the framework from systems that maximize global reference similarity without asking which attributes are being transferred.

3.5 Robust Inference and Calibration

Reference audio is often short, noisy, or mismatched in domain. A point estimate of z can therefore produce brittle personalization. The framework uses a perturbation-smoothed likelihood during training or adaptation so that the generator is stable in a neighborhood of the inferred reference factors. One form is

$$L_robust = E_\{z \sim q\varphi(z|R), \delta \sim N(0, \sigma^2 I)\}[-log\, p\theta(y \mid x, z + \delta)]$$

where qφ(z | R) is the variational reference posterior, δ is a Gaussian perturbation with variance σ², and I is the identity matrix. This expression encourages the generator to produce plausible outputs under small changes in the inferred reference representation. It is especially relevant for singing, where a short prompt may not reveal the performer's behavior across pitch registers, and for speech, where channel noise can dominate a brief reference.

Calibration complements robustness. A reference encoder can assign high confidence to a factor even when the prompt is ambiguous. Temperature scaling or posterior regularization can reduce overconfidence, and calibration diagnostics should be treated as part of model assessment rather than as an afterthought [21]. The key design rule is that uncertainty in z should propagate to synthesis: when the reference posterior is broad, the model should produce a distribution of plausible personalizations rather than an arbitrarily sharp output. This rule is compatible with the marginal likelihood in Section 3.1 and with the robust objective above.

4. Analytical Assessment and Design Trade-offs

4.1 Fidelity, Personalization, and Leakage

The framework can be assessed analytically through three quantities: content fidelity, personalization strength, and leakage. Content fidelity measures whether the generated acoustic sequence realizes the symbolic target; personalization strength measures whether requested attributes of the reference are reflected; leakage measures whether unrequested reference content, noise, or channel properties appear. A design criterion can be written as

$$J = Fidelity + \alpha\, Personalization - \beta\, Leakage$$

where α and β are nonnegative weights that encode task priorities. For read speech, content fidelity and speaker similarity may dominate; for singing, style and pitch-dependent timbre may receive more weight; for instrumental music, instrument identity and articulation may matter more than speaker-like identity. The expression is not a metric to be estimated from hidden data; it is a decision-theoretic summary of the trade-off that the training objective and inference procedure should respect.

The trade-off is structurally unavoidable. A reference embedding that is rich enough to capture expressive timing will also capture some content-correlated variation. A strong mutual-information penalty can remove that variation, but it can also remove legitimate correlations between style and timing. The framework handles this by assigning different regularization strengths to different subspaces: identity and timbre factors can be globally pooled and strongly regularized, while prosody factors remain temporal and more weakly regularized. This asymmetry reflects the physical structure of speech and music better than a single penalty applied uniformly to all factors.

4.2 Domain Differences Between Music and Speech

Speech and music share acoustic representations but differ in the causal structure of personalization. In speech, speaker identity is relatively stable across an utterance, while prosody varies with syntax, emphasis, and emotion. In singing, identity and timbre interact with pitch and vocal effort, and prosody is constrained by melody and rhythm. Instrumental music replaces speaker identity with instrument identity and performance style, and timbre is often the primary personalization target. These differences imply that a single reference encoder should not impose the same pooling or invariance across domains. Frame-level prosody encoders are appropriate for speech and singing, but a global timbre or instrument encoder is more appropriate for stable identity. A domain indicator can gate these pathways, but the gating should be trained with domain-specific content-adversarial objectives so that singing lyrics do not leak through a timbre pathway.

The music case also exposes a limitation of text-only content control. A musical score specifies pitch and duration, but articulation, dynamics, and micro-timing carry much of the expressive information. The reference style factor must therefore be allowed to influence timing without overwriting score content. The speech case exposes a different limitation: phoneme sequences do not fully specify rhythm, and reference prosody can legitimately change duration. The content-alignment loss must therefore be designed to permit controlled timing variation while penalizing lexical or melodic copying. These are not symmetric problems, and the framework's factor-specific conditioning routes are intended to make the asymmetry explicit.

4.3 Failure Modes and Boundary Conditions

Several failure modes follow from the formulation. First, posterior collapse can make z_s or z_t uninformative, in which case personalization vanishes while reconstruction remains acceptable. This is more likely when the reference encoder is over-regularized or when the decoder ignores conditioning. Second, leakage can persist through attention even when the global factors are clean, because attention weights can select content-bearing frames. Third, domain mismatch can produce a reference posterior that is broad but biased toward the source domain, leading to inappropriate timbre or prosody transfer. Fourth, adversarial training can sharpen outputs while amplifying artifacts under noisy references. Fifth, calibration failure can make the model appear confident about a personalization that is not supported by the reference.

The framework's boundary conditions are therefore clear. It is most defensible when the reference is sufficiently long to estimate the requested factors, when content and personalization can be separated by available weak supervision or architectural constraints, and when the evaluation distinguishes requested from unrequested transfer. It is less defensible when the reference is extremely short, when the target and reference share content that the model could copy, or when the personalization target is defined only by global similarity. In those cases, the factorization should be treated as a hypothesis about the data-generating process, not as a guarantee of causal separation.

5. Technical Discussion and Implications

5.1 Evaluation Logic Without Aggregate Similarity

Evaluation of reference-guided generation should not rely on a single similarity score between reference and output. Such a score conflates content fidelity, personalization, leakage, and realism. A more informative logic uses controlled perturbations: hold content fixed and vary style, timbre, or speaker; hold personalization fixed and vary content; introduce

distracting content or noise in the reference; and test whether the model transfers only the requested factor. These tests can be specified as analytical contrasts even before empirical validation. They also connect to the framework's objective: a model that succeeds should show stable content fidelity under style perturbation and stable style transfer under content perturbation, while leakage measures should increase when the reference contains copied content.

This evaluation logic also clarifies why calibration matters. If the reference posterior is broad and the model produces one confident output, a similarity score may reward a lucky match while hiding brittleness. Reporting uncertainty over generated personalizations, or evaluating the spread of outputs under reference perturbation, is more consistent with the marginal likelihood and robust objective. The framework therefore implies that evaluation should include sensitivity to reference length, noise, and domain shift, not only average quality on matched pairs.

5.2 Deployment Constraints and Ethical Use

Personalized acoustic generation raises legitimate concerns about identity misuse, voice cloning, and style appropriation. The factorization proposed here does not solve those concerns, but it makes some safeguards more natural. A deployment system can restrict which subspaces are available: timbre or speaker identity may require consent verification, while prosody and style may be allowed under weaker constraints. A content-adversarial pathway can reduce the chance that a reference supplies lexical or melodic content. Calibration and uncertainty reporting can flag weak consent signals or ambiguous references. These are governance choices, but they depend on the technical separation between requested personalization and nuisance transfer that the framework makes explicit.

Computationally, the framework is compatible with standard encoder-decoder synthesis. The reference encoder can use self-supervised or compression-style front ends [18], [19], [20], and discrete codes can reduce the cost of long reference modeling [17]. Cross-attention and modulation add parameters but avoid the need for a separate style classifier at inference. The main computational burden is the mutual-information estimator and adversarial training, which can be stabilized with contrastive bounds and Wasserstein objectives [24]. The design is therefore not a departure from existing synthesis architectures; it is a reorganization of their conditioning path around identifiable and regularized reference factors.

5.3 Implications for Representation Learning

The broader implication is that reference guidance is a representation-learning problem with a decision layer. The model must decide which reference factors are relevant, how confident it is about them, and how strongly to apply them to a target with different content. This view connects speech and music synthesis to disentanglement research [22], calibration [21], and content alignment [23], [25]. It also explains why simply enlarging the reference encoder may not improve personalization: a larger encoder can capture more detail, but detail is useful only if the conditioning path can select the right factors and reject leakage. The proposed factorization is a way to make that selection explicit and to expose the trade-offs to analysis.

6. Conclusion

This paper developed a reference-guided acoustic representation learning framework for personalized music and speech generation. The framework formalizes generation as a

marginal likelihood over reference factors, factorizes reference audio into style, prosody, timbre, and speaker or instrument subspaces, and conditions symbolic content through cross-attention and feature-wise modulation. It controls content leakage with mutual-information and content-adversarial terms, combines reconstruction, adversarial realism, content alignment, and posterior regularization in a single objective, and introduces perturbation-smoothed likelihood for reference uncertainty. Analytical assessment showed that fidelity, personalization, and leakage form a structural trade-off; that unsupervised disentanglement is insufficient without inductive bias; and that speech, singing, and instrumental music require different regularization of identity, timbre, and prosody. The framework provides a technically coherent basis for evaluation based on controlled perturbations and for deployment safeguards that distinguish requested personalization from unrequested transfer. Its significance lies in treating reference guidance as a representation and decision problem rather than as a single-vector conditioning trick, thereby making personalization more interpretable, regularized, and accountable.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 27 (pp. 2672–2680).

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

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998–6008).

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., Ying, C., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech (pp. 4006–4010).

6. 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).

7. 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.

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).

9. 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. 5180–5189).

10. Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

11. 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).

12. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer. arXiv preprint arXiv:1809.04281.

13. Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2018). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

14. Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341.

15. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. In Proceedings of the 34th International Conference on Machine Learning (pp. 1068–1077).

16. Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., & Cox, D. (2020). Unsupervised speech decomposition via triple information bottleneck. In Proceedings of the 37th International Conference on Machine Learning (pp. 7836–7846).

17. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In Advances in Neural Information Processing Systems 30 (pp. 6306–6315).

18. Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. arXiv preprint arXiv:2210.13438.

19. 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 33 (pp. 12449–12460).

20. 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.

21. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

22. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the 36th International Conference on Machine Learning (pp. 4114–4124).

23. Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (pp. 369–376).

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

25. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In Advances in Neural Information Processing Systems 28 (pp. 577–585).