Zero-Shot Multilingual Singing Voice Conversion Using Disentangled Speaker and Style Representations

Alister L. Neelin
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
neelin954@binghamton.edu

Salvador Sacristan
Department of Computer Science, University of Houston, Houston, TX, USA.
sacristan685@uh.edu

Annemarie Kostolzin
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
kostolzin877@oregonstate.edu

Qushi L. Bhattacharyya
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
qushi.bhattacharyya@ku.edu

Abstract

Zero-shot singing voice conversion aims to transform a source singing recording into the voice of an unseen target singer while preserving melodic, linguistic, and expressive content. The problem becomes substantially harder in multilingual settings, where source lyrics, target reference material, and singing conventions may differ across languages. A central obstacle is that speaker identity and singing style are statistically entangled: a short target reference may contain vibrato, dynamics, register, pronunciation habits, and recording conditions that are not intrinsic to the singer. This article develops a methodological framework for zero-shot multilingual singing voice conversion that factorizes a singing signal into language-agnostic content, speaker identity, singing style, and melody. The framework combines self-supervised content encoding, a reference-conditioned speaker encoder, a conditional style encoder, adversarial and correlation-based disentanglement penalties, and explicit melody preservation. Training objectives are formulated for reconstruction, content consistency, melody accuracy, and speaker-style independence, and an inference rule is given for controlling style transfer through latent interpolation. The analysis addresses cross-lingual generalization, style-speaker confounding, pitch-range mismatch, error propagation through neural vocoding, and evaluation logic. The contribution is a technically explicit design and analytical assessment that connects disentangled representation learning to the practical requirements of zero-shot multilingual singing voice conversion.

Keywords

zero-shot singing voice conversion; speaker disentanglement; style representation; multilingual speech; self-supervised learning; neural audio synthesis

1. Introduction

Singing voice conversion changes the perceived singer identity of a recording while retaining the linguistic content and musical contour of the source performance. Unlike conversational voice conversion, singing voice conversion must preserve not only phonemes but also sustained pitch, vibrato, timing, dynamics, and expressive articulation. These factors are closely coupled with singer identity: a target singer's vocal tract and glottal characteristics shape timbre, while the singer's habitual use of vibrato, register transitions, consonant articulation, and rhythmic inflection shapes style. Consequently, a converter that accepts a short reference recording must decide which observed factors belong to the target speaker and which belong to the target performance context.

Early one-shot and zero-shot voice conversion systems demonstrated that speaker and content factors can be separated through instance normalization, autoencoding, adversarial training, and cycle consistency [1], [2], [3], [4]. Self-supervised speech representations further improved content encoding by providing robust frame-level features learned from large unlabeled corpora [5], [6]. Neural audio compression and analysis-synthesis systems then made it possible to reconstruct high-fidelity waveforms from compact latent representations [7], [10]. In parallel, speech synthesis research showed that prosody and style can be represented as latent variables or reference-derived embeddings and transferred independently of linguistic content [8], [9]. These developments provide the representational ingredients for zero-shot singing voice conversion, but they do not resolve the specific interaction between singer identity and singing style.

Singing-specific methods have addressed pitch preservation, spectral detail, and expressive control. Direct waveform modification and statistical singing voice conversion established the importance of pitch and spectral envelope modeling in non-parallel settings [12]. Generative adversarial approaches introduced flexible many-to-many mappings [13], and shared challenge protocols clarified evaluation conditions for cross-lingual voice conversion [14]. Recent work on zero-shot singing voice synthesis with style transfer and multi-level style control shows that singing systems can be conditioned on reference style information while generating unseen singers [11]. The present article is not a report of a particular system evaluation; rather, it formulates a coherent methodological framework for zero-shot multilingual singing voice conversion built around disentangled speaker and style representations. The key claim is that robust zero-shot conversion requires an explicit factorization of content, speaker identity, style, and melody, together with training constraints that discourage the style latent from carrying speaker identity and the speaker latent from carrying expressive performance traits.

The article makes three contributions. First, it defines a factorized latent model for multilingual singing voice conversion in which content is language-agnostic, speaker identity is reference-conditioned, style is separable from identity, and melody is preserved as an explicit control signal. Second, it develops a set of training objectives that combine reconstruction, content consistency, adversarial disentanglement, correlation penalties, and melody preservation. Third, it provides an analytical assessment of cross-lingual generalization, style-speaker confounding, pitch-range mismatch, and vocoder-induced error, and it specifies evaluation logic that follows from the proposed factorization. The discussion is intentionally grounded in the mechanics of representation learning and signal reconstruction rather than in unsupported empirical claims.

2. Problem Formulation and Representation Decomposition

Let x denote a source singing signal with linguistic content and melody, and let r denote a short reference recording of an unseen target singer. The converter produces \hat{x}, which should preserve the source content and melodic contour while adopting the target singer's timbre. In multilingual operation, x and r may come from different languages and may follow different singing traditions. The output retains the source lyrics unless an explicit language-adaptation control is applied; therefore, the converter must avoid treating language identity as part of speaker identity or style.

The proposed representation model factorizes the singing signal into content z_c, speaker identity s, style y, and melody m. The content variable should encode phonetic and syllabic structure in a form that is stable across languages. The speaker variable should encode vocal-tract and glottal characteristics that persist across songs and languages. The style variable should encode expressive and performance-level factors such as vibrato rate and depth, dynamics, phonation type, articulation, rhythmic inflection, and expressive timing. The melody variable should encode the pitch contour and its first-order variation, which are essential for singing and must not be overwritten by speaker conversion. The generative relation is expressed as follows.

$$x \approx G(z_c, s, y, m), z_c = E_c(x), s = E_s(r), y = E_y(r)$$

Here, G is the conversion and synthesis network, E_c is the content encoder, E_s is the speaker encoder, and E_y is the style encoder. The symbol ≈ indicates approximate reconstruction under the model; it does not assert exact invertibility. The factorization is useful only if the variables are constrained to remain functionally distinct. In particular, z_c should not predict s or y, s should not predict y, and m should remain recoverable from the converted output. These constraints define the disentanglement problem.

A second issue concerns zero-shot operation. Because r is not paired with training targets for the same singer, s must generalize from a short reference. If r contains a distinctive style, E_s may absorb style information, producing a converted voice that imitates the reference performance rather than the underlying singer. Conversely, if E_y absorbs timbre, style transfer will alter identity. The framework therefore treats disentanglement as a first-class modeling constraint rather than an incidental property of the encoder architecture.

3. Proposed Zero-Shot Multilingual Framework

3.1 Content and Language Encoding

The content encoder maps the source signal to frame-level representations that capture phonetic identity, syllabic organization, and coarse timing. Self-supervised speech models of the HuBERT and wav2vec 2.0 families provide contextual representations that are robust to speaker variation and recording conditions [5], [6]. For multilingual singing, the encoder should be initialized or adapted with multilingual material so that its feature space is not tied to a single phoneme inventory. Language-adversarial training can further reduce language-specific leakage from z_c when the available languages are unevenly represented. The encoder does not need to transcribe lyrics explicitly; however, an auxiliary phonetic posterior or self-supervised masked-prediction objective can stabilize content learning during training.

Because singing stretches syllables and introduces long stable vowels, frame-level content features must preserve temporal alignment. Self-attention and convolutional downsampling can model local phonetic context while retaining the temporal resolution required for vibrato

and note transitions [15]. The content variable is therefore not a single utterance embedding but a sequence z_c = (z_{c,1}, ..., z_{c,T'}), where T' is the encoded sequence length. This sequence is later aligned with melody and style during conversion.

3.2 Speaker and Style Encoders

The speaker encoder E_s maps the target reference r to a vector s ∈ R^{d_s}. The style encoder E_y maps the same or another reference to y ∈ R^{d_y}. A conditional variational formulation allows the style encoder to represent variability in expressive performance while keeping the speaker vector deterministic [16]. The style posterior can be conditioned on content and melody so that the latent variable explains expressive variation rather than phonetic identity. The speaker vector is trained to be stable across content, pitch, and language, while the style vector is trained to explain performance-level variation that may change within a single singer.

Adversarial classifiers provide a practical constraint. A speaker classifier D_s attempts to predict singer identity from the style latent, and a style classifier D_y attempts to predict style from the speaker latent. The encoders are trained to reduce the success of these classifiers, while the classifiers are trained to remain accurate. This min-max relation discourages style latents from carrying identity and speaker latents from carrying expressive style.

$$L_adv = \max_{D_s}\min_{E_y}\mathrm{E}\left[\log D_s\left(s|z_y\right)\right] + \max_{D_y}\min_{E_s}\mathrm{E}\left[\log D_y(y|z_s)\right]$$

In this expression, D_s(s|z_y) is the predicted probability of the true speaker label given the style latent, and D_y(y|z_s) is the predicted probability of the style label given the speaker latent. The expectation is taken over training examples and their associated labels. The min-max ordering states that each classifier is optimized for accuracy while the corresponding encoder is optimized to make its latent uninformative about the other factor. This is a training principle, not evidence that disentanglement is perfect; residual leakage is a recognized limitation of adversarial factorization.

3.3 Melody and Prosody Preservation

Melody is not a minor attribute in singing conversion. Pitch contour carries the musical identity of the performance, and errors in F0 often sound more severe than minor timbral deviations. The melody extractor M estimates a pitch contour and its first difference from the source or from an explicit musical score. The contour is normalized only for range adaptation, not for expressive shape. The converter receives m as a control signal and is trained to reproduce both the contour and its local variation. Energy and duration can be handled as auxiliary prosodic controls, but they should not be conflated with speaker identity.

Disentangling melody from speaker is necessary because a target singer may have a different comfortable register. Directly copying the source F0 can produce unnatural phonation, while over-normalizing to the target range can erase expressive pitch gestures. A practical design separates note-level pitch targets from within-note expressive movement. The converter may shift the overall register to match the target range while preserving interval structure and vibrato shape. This separation is especially important in multilingual singing, where ornamental pitch gestures may be language- or tradition-specific.

3.4 Conversion and Neural Synthesis

The converter G combines content, speaker, style, and melody. A typical architecture uses content-melody alignment, speaker conditioning through adaptive normalization, and style

conditioning through attention or feature-wise modulation. The output is a spectral representation or a neural codec latent that is subsequently decoded to a waveform. Neural waveform generators of the Parallel WaveNet family are a natural synthesis component because they support high-fidelity parallel generation [18]. The vocoder should be trained or adapted on singing material because sustained vowels and wide dynamic range expose artifacts that are less prominent in speech.

At inference, the target speaker vector is extracted from the unseen reference r_t. The style vector may be taken from the source performance, from the target reference, or from an interpolation between the two. This choice determines whether the system preserves the source singer's expressive style while changing identity or imitates the target reference's style. The framework therefore exposes style as a controllable factor rather than forcing it to be copied implicitly with timbre.

4. Training Objectives and Inference

4.1 Reconstruction and Content Consistency

The primary reconstruction objective requires the converted output to match the source signal when the source speaker, style, and melody are used as conditions. A spectral reconstruction term reduces over-smoothing, while a waveform or codec-domain term preserves phase and transient detail. Let Φ denote a differentiable spectral magnitude transform. The reconstruction loss is written as follows.

$$L_rec = E[||x - G(z_c, s, y, m)||_1 + ||\Phi(x) - \Phi(G(z_c, s, y, m))||_2\text{^}2]$$

The first term encourages waveform or latent fidelity, and the second term emphasizes spectral envelope and harmonic structure. The expectation is over source-reference pairs used during training. The symbol s here denotes the source speaker vector during reconstruction; at conversion time it is replaced by the target speaker vector.

Content consistency is enforced by re-encoding the converted output and requiring its content representation to match the source representation. This constraint is particularly important in multilingual training, where the converter might otherwise suppress phonemes that are rare in the target language.

$$L_con = E[||E_c(G(z_c, s, y, m)) - z_c||_2\text{^}2]$$

Here, E_c is the same content encoder used for the source, and the loss is evaluated on the converted output. The term does not require parallel singing data; it requires only that the content encoder remain stable under conversion. In practice, a stop-gradient or slowly updated target encoder can reduce instability during adversarial training.

4.2 Speaker-Style Disentanglement

The adversarial loss in Section 3.2 is complemented by a correlation penalty that discourages pairwise dependence among content, speaker, and style latents. Exact statistical independence is difficult to estimate in high dimensions, so a covariance-based penalty provides a stable surrogate. Let z_a and z_b denote two latent variables from the set {z_c, s, y}. The penalty is as follows.

$$L_corr = \Sigma_\{a \neq b\} \, ||Cov(z_a, z_b)||_F\text{^}2$$

The operator Cov computes the empirical cross-covariance between latent dimensions, and ||·||_F is the Frobenius norm. The summation runs over ordered pairs of distinct variables. This term penalizes linear dependence, while the adversarial term targets nonlinear

predictability. Neither term guarantees identifiability, but together they reduce the most common failure mode in short-reference singing conversion: the style encoder learning a singer-specific timbre signature. A beta-VAE regularization on the style posterior further encourages disentangled and smooth style factors [19], and Wasserstein adversarial training can stabilize the min-max optimization [20].

4.3 Melody Loss and Total Objective

Melody preservation is trained directly on the pitch contour and its first difference. Let M be a differentiable pitch extractor, and let Δ denote the first temporal difference. The melody loss is written as follows.

$$L_mel = E[||M(G(z_c, s, y, m)) - m||_1 + ||\Delta M(G(z_c, s, y, m)) - \Delta m||_1]$$

The first term preserves the absolute pitch contour after any permitted register shift, and the second term preserves local pitch movement such as vibrato and portamento. The total training objective combines the terms with nonnegative weights selected to balance reconstruction quality against disentanglement and melodic fidelity.

$$L = \lambda_r ecL_r ec + \lambda_c onL_c on + \lambda_c orrL_c orr + \lambda_a dvL_a dv + \lambda_m elL_m el$$

The weights λ_rec, λ_con, λ_corr, λ_adv, and λ_mel control the trade-off between fidelity and factorization. Increasing λ_corr and λ_adv reduces leakage but may discard expressive information that is genuinely correlated with singer identity. Increasing λ_mel preserves pitch but can restrict natural range adaptation. These trade-offs are central to system design and should be assessed through controlled ablations rather than assumed.

4.4 Zero-Shot Inference and Style Control

At inference, the target speaker vector is s_t = E_s(r_t), where r_t is the unseen singer's reference. The source content and melody are extracted from x. The output style can be selected from the source performance y_x = E_y(x), from the target reference y_t = E_y(r_t), or from a convex combination. A normalized interpolation is used to keep the style latent within the region observed during training.

$$y_out = ((1-\alpha)y_x + \alpha\, y_t) \,/\, ||(1-\alpha)y_x + \alpha\, y_t||_2$$

The coefficient $\alpha \in [0,1]$ controls the balance between source style preservation and target-reference style imitation. When α = 0, the converted voice keeps the source performance style while adopting the target timbre. When α = 1, the target reference style is transferred more strongly. This inference rule makes the speaker-style distinction operational: identity comes from s_t, while expressive style is set by y_out. The converter then generates \hat{x} = G(z_c, s_t, y_out, m). A language control may be added when the target reference is in a different language, but it should be conditioned separately from speaker and style to avoid language leakage.

5. Analytical Assessment and Failure Modes

5.1 Cross-Lingual Generalization

Multilingual singing conversion stresses the content encoder because phoneme inventories, syllable structures, and singing articulation differ across languages. Self-supervised representations are useful because they are learned from broad acoustic distributions rather than from a fixed phoneme set [5], [6]. Nevertheless, these representations can retain language-specific regularities, especially when pretraining data are dominated by

one language. A converter trained mostly on high-resource languages may map low-resource phonemes to nearby high-resource categories, producing intelligibility loss that is not reflected in speaker similarity. Language-adversarial training and balanced multilingual sampling reduce this risk, but they cannot compensate for phonemes that are absent from the training distribution. The evaluation should therefore separate in-language conversion from cross-language conversion and report content intelligibility by language pair.

A related failure mode is prosodic transfer. Singing traditions differ in ornamentation, vibrato, vowel modification, and rhythmic placement. If the style encoder is trained without language or tradition labels, it may learn a mixture of style and language identity. The output can then sound like a target singer performing in the wrong tradition even when the phonemes are correct. The proposed factorization addresses this only partially: language conditioning must remain separate from style conditioning, and the style latent should be regularized toward smooth, low-dimensional variation rather than allowed to encode language identity.

5.2 Style-Speaker Confounding

The most difficult theoretical issue is non-identifiability. Given only observed singing, multiple factorizations can explain the same data. A style encoder may encode breathiness that is actually caused by the singer's vocal physiology, and a speaker encoder may encode vibrato that is actually a stylistic choice. Adversarial classifiers and covariance penalties reduce linear and nonlinear predictability, but they do not prove that the true generative factors have been recovered. The practical consequence is that zero-shot conversion can sound convincing while transferring the wrong attribute. For example, a target reference with a breathy, intimate performance may cause the converted output to become breathy even when the requested style is neutral. The style interpolation in Section 4.4 mitigates this by allowing the source style to dominate, but the speaker vector may still contain residual style information.

This limitation argues for explicit inductive biases: separate pitch and energy normalization before speaker encoding, style labels or disentanglement constraints where available, and reference augmentation that varies style while holding singer identity constant. Such augmentation is not a substitute for data coverage, but it makes the desired invariance explicit during training.

5.3 Pitch-Range and Vocoder Error Propagation

Pitch handling introduces a second trade-off. If the source F0 is copied exactly, the target singer may be forced outside a natural register, causing strain or octave errors. If the contour is strongly normalized, expressive pitch gestures may be flattened. Separating note-level pitch from within-note movement allows a register shift that preserves intervals and vibrato shape. The melody loss in Section 4.3 encourages this behavior, but it must be paired with a range constraint so that the converter does not satisfy the loss by producing unnatural phonation.

Neural vocoding adds another error source. Singing has long sustained vowels, wide dynamic range, and pronounced harmonic structure. A vocoder trained primarily on speech may introduce buzzy artifacts, unstable vibrato, or harmonic discontinuities. These errors are not captured by frame-level content or speaker losses. Evaluation should therefore include signal-level measures of harmonicity, spectral envelope fidelity, and artifact detection, alongside perceptual naturalness. The converter should not be judged only by speaker similarity, because a high speaker score can coexist with unstable pitch and poor intelligibility.

5.4 Evaluation Logic

The proposed factorization suggests an evaluation protocol with four separable axes: speaker identity, content intelligibility, melody preservation, and style fidelity. Speaker identity should be measured with an unseen-speaker verification model and with human judgments on converted voices from multiple source songs. Content intelligibility should be evaluated with automatic speech recognition and human transcription, stratified by language pair. Melody preservation should compare pitch contours and note transitions, not only average F0. Style fidelity should assess vibrato, dynamics, phonation, and articulation using both reference-based similarity and attribute ratings. Naturalness and artifact ratings should be reported separately. This separation follows directly from the model: if the factors are meant to be independently controllable, aggregate scores that mix them cannot diagnose leakage.

6. Discussion and Implications

The framework has practical implications for multilingual singing applications. A zero-shot converter with a disentangled speaker representation can adapt to a new singer from a short reference, which reduces the need for studio recordings and parallel data. A separate style representation supports expressive control, such as preserving the source performance's intimacy while adopting the target singer's timbre. In cross-lingual settings, language-independent content encoding helps prevent the converter from collapsing unfamiliar phonemes into the target language's inventory. These capabilities are relevant to multilingual covers, dubbing, and creative tools, provided that consent, attribution, and misuse safeguards are addressed.

The main scientific implication is that zero-shot singing voice conversion is not only a mapping problem but also an invariance problem. The model must learn what should change, namely singer identity, and what should remain stable, namely content, melody, and selected style factors. Adversarial and correlation penalties are useful because they make these invariances explicit, but they are not sufficient on their own. Robustness depends on data coverage, reference augmentation, pitch-range modeling, and vocoder quality. The framework therefore treats disentanglement as a design objective with measurable failure modes rather than as a property guaranteed by architecture.

7. Conclusion

This article formulated a methodological framework for zero-shot multilingual singing voice conversion using disentangled speaker and style representations. The framework factorizes singing into language-agnostic content, speaker identity, style, and melody, and it specifies encoders, adversarial and correlation-based disentanglement constraints, reconstruction and content-consistency losses, melody preservation, and an inference rule for style control. The analysis identified cross-lingual content leakage, style-speaker confounding, pitch-range mismatch, and vocoder error as principal failure mechanisms. It also specified an evaluation logic that assesses identity, content, melody, style, and naturalness separately. The central conclusion is that zero-shot multilingual singing conversion requires explicit invariance constraints and separate handling of melody and style; without them, a short target reference can entangle identity with performance context and undermine both speaker fidelity and cross-lingual intelligibility. The proposed formulation provides a technically coherent basis for system design and for diagnostic evaluation of disentangled singing voice conversion.

References

1. Chou, J.-C., Yeh, C.-C., & Lee, H.-Y. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. In Proceedings of Interspeech 2019 (pp. 664–668).

2. Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 266–273). IEEE.

3. Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). AutoVC: Zero-shot voice style transfer with only autoencoder loss. In Proceedings of the 36th International Conference on Machine Learning (pp. 5210–5219). PMLR.

4. Kaneko, T., & Kameoka, H. (2018). CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2100–2104). IEEE.

5. 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.

6. 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).

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

8. Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., 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. 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). PMLR.

10. Choi, H., Lee, J., Kim, W., Lee, J., Heo, H., & Lee, K. (2021). Neural analysis and synthesis: Reconstructing speech from self-supervised representations. In Advances in Neural Information Processing Systems (Vol. 34, pp. 16251–16265).

11. 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).

12. Kobayashi, K., Toda, T., Neubig, G., Sakti, S., & Nakamura, S. (2014). Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In Proceedings of Interspeech 2014 (pp. 2514–2518).

13. Sisman, B., Zhang, M., Dong, M., & Li, H. (2020). On the study of generative adversarial networks for singing voice conversion. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7234–7238). IEEE.

14. Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., & Toda, T. (2020). Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In Proceedings of the Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020 (pp. 80–98).

15. 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 (Vol. 30, pp. 5998–6008).

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

17. 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 (Vol. 27, pp. 2672–2680).

18. van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Sifre, L., & Hassabis, D. (2018). Parallel WaveNet: Fast high-fidelity speech synthesis. In Proceedings of the 35th International Conference on Machine Learning (pp. 3918–3926). PMLR.

19. Higgins, I., Matthey, L., Pal, A., Burgess, C., 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.

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