A COMPARATIVE ANALYSIS OF LATENT REGRESSOR LOSSES FOR SINGING VOICE CONVERSION

Brendan O'Connor
Media Arts Technology CDT,
Centre For Digital Music,
Queen Mary University of London
b.d.oconnor@qmul.ac.uk

Simon Dixon
Centre For Digital Music,
Queen Mary University of London
s.e.dixon@qmul.ac.uk

ABSTRACT

Previous research has shown that established techniques for spoken voice conversion do not perform as well when applied to singing voice conversion (SVC). We propose an alternative loss component in a loss function that is otherwise well-established among VC tasks, which has been shown to improve our model's SVC performance. We first trained a singer identity embedding (SIE) network on mel-spectrograms of singer recordings to produce singer-specific variance encodings using contrastive learning. We subsequently trained a well-known autoencoder framework (AutoVC) conditioned on these SIEs, and measured differences in SVC performance when using different latent regressor loss components. We found that using this loss w.r.t. SIEs leads to better performance than w.r.t. bottleneck embeddings, where converted audio is more natural and specific towards target singers. The inclusion of this loss component has the advantage of explicitly forcing the network to reconstruct with timbral similarity, and also negates the effect of poor disentanglement in AutoVC's bottleneck embeddings. We demonstrate peculiar diversity between computational and human evaluations on singer-converted audio clips, which highlights the necessity of both. We also propose a pitch-matching mechanism between source and target singers to ensure these evaluations are not influenced by differences in pitch register.

1. INTRODUCTION AND BACKGROUND

As domains like speech and computer vision produce more convincing and novel probabilistic machine learning-based transformations, the inevitable interest of the entertainment industry has propelled further research in more artistic-relevant domains such as singing. The ability to switch between multiple singers without the target singer needing to re-record songs will be revolutionary to the music industry in the amount of time and money it saves, not to mention the artistic avenues this opens for up for composers and music producers. This is made possible by neural networks that can achieve singing voice conversion (SVC),

which is the task of changing the perceived identity of a singer in an audio recording. The majority of voice conversion research has focused on the speech domain, while the singing domain has only gained attention in recent years.

To achieve voice conversion, a typical system comprises an encoder that produces vocalist-independent embeddings, usually representing linguistic content from input data. A decoder then combines these embeddings with the speaker-specific timbre representations on which it was conditioned to resynthesise the data in its voice-converted state.

The conditioning factor of speaker timbre can be represented in a one-hot encoding format. However, this is inherently only able to train a network for conversions to the finite number of vocalists seen in the training data [1–3]. More recent research is focused on zero-shot conversions, where systems are able to take unseen examples as both the source and target signal. AutoVC [4] is a well known voice conversion network consisting of the conditioned decoder system described above, and can achieve zero-shot conversions by replacing one-hot conditioning vectors with 'voice identity embeddings' which contain vocalist-specific variances. Other conditioning factors used to disentangle timbre include loudness [1,5], phonetic (usually from pretrained linguistic networks) [1,5–9], and pitch-related features either as one-hot encodings or continuous data [5, 8–12].

Voice identity embeddings are often averaged across multiple speaker utterances in order to get a generalised representation of how a given speaker sounds in different conditions. [11] demonstrated better voice conversion performance when using an additional network that takes F0 information, live-generated and speaker-averaged embeddings to produce a new *adjusted* embedding. [9] proposed a U-net to model with instance normalisation modules [13] after each downsampling block to produce a hierarchical speaker embedding at multiple granularities.

Conditioning with these embeddings does not however, imply perfect disentanglement from the linguistic information. Methods towards improving disentanglement in these systems include fine-tuning the size of the bottleneck to have a capacity capable of retaining only linguistic content [4, 11], or applying a vocalist-classifier to the bottleneck and using its negative loss as a regularisation component in the autoencoder loss function (which have also been used on converted data for training and evaluation) [3].

Variational autoencoders (VAEs) alone have been used for voice conversion [2], while VAEs combined with GAN frameworks during the training phase have been shown to perform timbre conversion between voice and other instruments [14]. GANs have also been applied for voice conversion where the generator is either a fully convolutional network [15], VAE [16, 17] encoder-decoder framework with cycle-consistency [18], or U-net [19] framework.

When feature-matching [20–22] is implemented between original and reconstructed data in the latent space, this can be referred to as the latent regressor loss or cycle-consistency [4, 5, 23–25]. In most of the literature, cycle-consistency more often refers to the use of back-translation [26] in generative networks to complete a two-step conversion cycle of source to target to source [18, 25, 27, 28]. As the implementation in [4] describes something more similar to a latent regressor loss [29], we will herein use this terminology. Latent loss provides the additional intuition that generative networks must rely on the relevant latent space embeddings in addition to other inputs. Loss functions that penalise silences and attribute heavier weightings towards spectral peaks have also been used, as well as multi-resolution losses [8]. Latent losses w.r.t. autoencoder bottleneck [4, 11, 23] and speaker embeddings [30, 31] have been used in speaker conversion models of varying frameworks. Recent work [32] demonstrated that training without a bottleneck latent loss accelerated model convergence for the task of singing technique conversion, producing audio with more detail and less audible artefacts.

In this paper, we train the well known voice conversion model AutoVC [4] and its voice identity encoder on singing data to perform SVC. We use different loss functions to reflect pixel-wise and higher level characteristic comparisons. Using objective metrics and subjective evaluations, we compare how these loss functions affect information disentanglement at AutoVC's bottleneck, the quality of its output audio clips, and the amount of perceived voice conversion these audio clips present.

2. EXPERIMENTS

2.1 Singing Voice Conversion System

Due to its influence on subsequent voice conversion research [5, 10, 10, 11, 33, 34], we have adopted the AutoVC architecture and explored adaptations to its loss functions. A flowchart of this system and these adaptations is provided for reference in Fig. 1. It consists of an autoencoder, where the encoder and decoder are conditioned on embeddings that represent variances specific to a given singer. These singer identity embeddings (SIEs) come from a pretrained network and can be conceptually summarised as representing singer-specific timbral qualities. For convenience, we will refer to this autoencoder section and the pretrained encoder as the SVC and SIE networks respectively (for greater detail on this architecture, readers can refer to the original AutoVC paper [4]).

The input and output data of the network is in a mel-spectrogram format. It is trained using the same input data for both SIE and SVC networks. The input features used

are 128 frames of 80 dimensional log mel-spectrograms of singer recordings, generated from 16kHz audio with an FFT size of 1024 and a hop size of 256 (frame duration of 16ms). After training, conversions can be achieved by feeding the SVC encoder with mel-spectrograms of a source singer, while the SIE network is fed those of a target singer. As the SVC decoder has been conditioned with SIEs, it is trained to utilise the target singer SIEs to reconstruct the input data at the conversion phase, provided that the bottleneck dimensions are not large enough to encode them. For this reason, a calibration process of the bottleneck capacity is required in order to disentangle timbre from the input data adequately. However, this empirical process can be time-consuming, qualitatively ambiguous and inherently causes a trade-off between good SVC and audio quality.

Our implementation of AutoVC incorporates the L1 loss to quantify reconstruction, and a secondary weighted loss which represents the difference between the SVC encoder embeddings for original data and reconstructed data, acting as a bottleneck latent regressor loss (BN-LR). To follow up with inconsistencies regarding the effect of BN-LR loss [32], we investigate whether the exclusion of this loss would be beneficial in the context of SVC. As an alternative solution, we also investigate how an SIE latent regressor (SIE-LR) loss would affect SVC, as we believe this may improve performance without the need for bottleneck capacity calibration. To summarise, the objective functions we will be comparing in our experiments include the L1 loss for: just reconstruction; reconstruction and bottleneck latent loss; and reconstruction and SIE latent loss, as defined in Equations 1, 2 and 3 respectively.

$$L_{RECON} = L_1(\hat{X}, X) \quad (1)$$

$$L_{BN-LR} = L_1(E_{SVC}(\hat{X}), E_{SVC}(X)) \quad (2)$$

$$L_{SIE-LR} = L_1(E_{SIE}(\hat{X}), E_{SIE}(X)) \quad (3)$$

As we are attempting to disentangle timbre from the input signal, we will refer to the bottleneck embeddings as the *non-timbral* information, which includes attributes independent to singer identity such as pitch contours, singing techniques, phonetics and other unaccounted variances. In the speech domain, there is a consideration of whether accent is included in timbral perception. However this is usually outside the scope of SVC research, likely due to very little accent diversity within datasets.

2.2 Datasets

We choose to use the Digital Archive of Musical Performances (DAMP) Intonation dataset [35] for this experiment. It is a subset of the singing data supplied by a publicly available karaoke app designed by Smule[1]. It consists of 4702 unprocessed recordings of 3556 singers, created in acoustically untreated environments primarily with mobile phone devices. This means there is inherently a

---

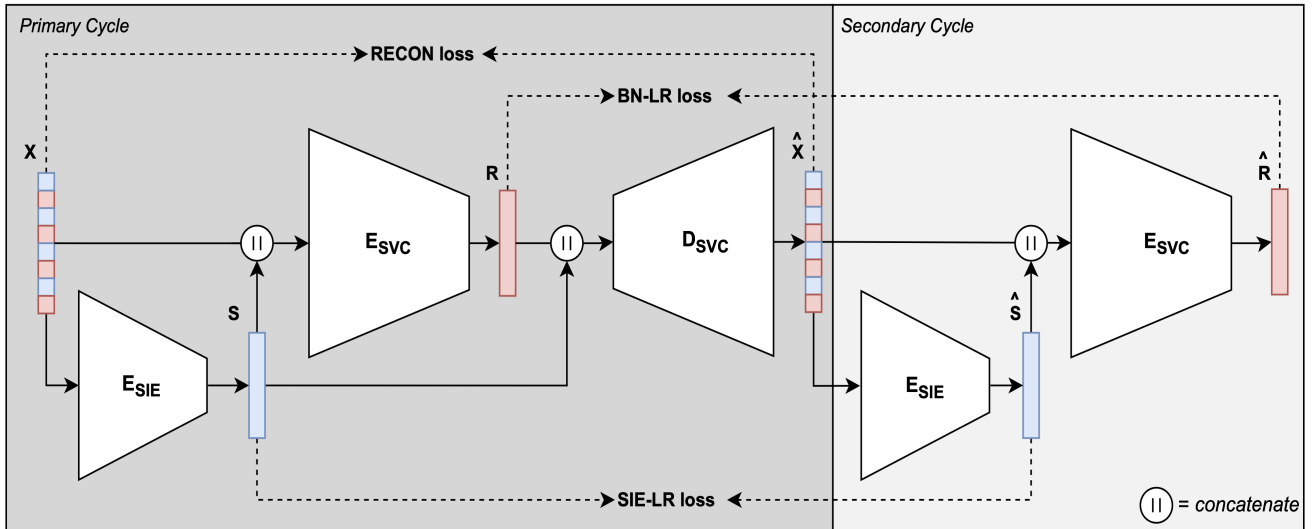[1] https://ccrma.stanford.edu/damp/

Figure 1. Diagram of the encoder/decoder components in the AutoVC network, with a secondary cycle partition illustrating how encodings for the reconstructed data are obtained. Vector comparisons used for reconstruction loss (RECON), bottleneck regressor loss (BN-LR) and singer identity embedding regressor loss (SIE-LR) are shown with dotted connectors. X represents a mel-spectrogram containing timbre and non-timbral content, while S (blue) and R (red) represent vectors of these attributes respectively in their disentangled states.

considerable amount of background noise such as faintly heard backing tracks, ambience and other miscellaneous sound events. It also consists of many non-singing segments due to singers waiting for their backing tracks' instrumental sections to conclude. We remove these segments to the best of our abilities using an empirically selected volume threshold to detect vocal activity and removing chunks that remain below this threshold.

2.3 Training and Synthesis

To train the SIE network, we use a 3-layer LSTM network, trained using the GE2E loss objective [36]. This works as a self-supervised model, using a contrastive learning loss to encourage randomly selected recording clips of same singers to cluster in the output embedding's latent space, while repelling clusters of different singers. This was trained on the DAMP dataset with a batch size of 80 (8 singers * 10 utterance instances), with early stopping set to 40 iterations. The network was trained for a total of 314k iterations. The combination of the autoencoder and SIE network as described in Section 2.1 allows us to perform SVC. We preserved 80% of it as training data, while the remainder was split between validation and test subsets. We train an AutoVC model with a reconstruction loss only (RECON); reconstruction with bottleneck latent regressor loss (RECON+BN-LR); and reconstruction with SIE latent regressor loss (RECON+SIE-LR). These models are trained for a total of 500k iterations, using a batch size of 2 and the ADAM optimizer with a learning rate of 10^{-4}.

As this experiment investigates the effects of loss components on SVC, we are not making claims about SOTA results. We can therefore afford to reduce our dataset to 75% of its original size (keeping subset ratios the same) as used for SIE training, in the interest of saving on computation time. We reduced this further by using precal-

culated SIEs for each singer by running the SIE network sequentially across windowed chunks of all data for each singer. We then took the mean of these embeddings, resulting in an average SIE representation for each singer (as done to determine the GE2E loss used for SIE training [36]). The embeddings were encapsulated as a lookup table to be used during SVC training and synthesis to save on computational time.

In previous work on singing voice conversion [5], a similar architecture was trained with pitch-conditioning embeddings. During the conversion, pitch contours from the source data were transposed by an octave to match the octave closest to the average range of the target singer. We consider this to be a forceful application of pitch shifting that does not take the source singer's pitch range into consideration or how singer timbre can change with pitch - but in the context of some SVC methods or contexts we appreciate the necessity to do so. However, for complete authenticity in SVC, we chose not to apply octave shifts. Instead we applied a custom pitch-matching algorithm that chooses a target singer provided that there is an example of them singing roughly in the same range as the given source singer clip. This also ensures that evaluations of the network's performance can compare the speaker identities of two audio clips without being affected by changes in pitch register.

2.4 Performance Evaluation

To evaluate our model, we use a combination of objective metrics and subjective human evaluations. Relying on computational metrics alone can otherwise be misleading, as some methods of optimising a network may neglect important aspects of data structures. This can ultimately lead to poor reconstruction or conversions when evaluated by human listeners.

For subjective evaluations, we recruited 23 participants from MIR communities, ranging from age 24 to 53, to listen to a number of audio clips generated by the three aforementioned models under the following different source-target gender conditions: $M - M, M - F, F - M, F - F$, where M and F stand for male and female respectively. They were asked to rate how similar voices in these audio clips were to given target voices using a 5-point scale, as well as how natural the voices sounded.

To synthesise the audio from mel-spectrograms, we used a Wavenet vocoder provided by the authors of [4] [2], which was pretrained on the VCTK dataset [37] and conditioned on mel-spectrograms for spectrogram-to-audio conversion tasks. We expect that this will induce some degree of audio degredation, which should be accounted for. To do this, we converted the audio of four additional singer clips to mel-spectrograms which were resynthesized back (without passing through our SVC network) to audio using the Wavenet vocoder. These examples, along with 16 other voice-converted audio clips where evaluated by participants for audio quality and conversion performance. Each listener was given a unique set of audio clips, generated to satisfy each of the experimental conditions. A mean opinion score (MOS) across participants was then calculated for each condition.

As we considered the possibility that perception of naturalness and similarity could be correlated, cosine similarities between SIEs of converted audio and SIEs of the target singer were included in our analysis, which allowed us to investigate the models' conversion performance in a manner that is disentangled from naturalness.

To analyse the amount of disentanglement between timbral and non-timbral information, we applied a classification layer to pretrained AutoVC encoders, and trained this layer on 20 singers from our DAMP test subset. Resulting classification accuracies allow us to compare the amount of singer-specific timbre information remaining in AutoVC's bottleneck vectors between different models.

3. RESULTS

3.1 Voice Conversion and Naturalness

The top two graphs in Fig. 2 show the MOS results for perceptual naturalness, perceptual similarity and cosine similarity under different conditions, grouped by condition types with colour.

Note that the MOS naturalness rating for original audio clips was 3.72 (displayed in red at the top of Fig. 2 for reference), which can be considered as the approximate upperbound ceiling of perceptual evaluation due to the resynthesis process.

In previous work on singing technique conversion using similar methods [38], no significant correlation was found between similarity and naturalness. However, in this experiment a correlation between the two rating types is noticeably visible, and a Pearson test between the two ratings resulted in r=0.83, p < 0.002. Due to the strength of this positive correlation, we must be careful not to assume that

participant's perception of successful voice conversion are disentangled from their perception of naturalness. We hypothesise that while participants were asked to rate how confident they were that the reference recording and the voice-converted recordings were made by the same singer, lower similarity ratings may be caused by one or both of these two circumstances: artefacts caused by the SVC process in the audio file led to a lack of clarity in the timbre; or the timbre of the converted voice itself was indeed dissimilar to the target voice.

Cosine similarities are shown in the bottom chart of Fig. 2. These scores reinforce the perceptual similarity scores for the different model conditions. We can see that compared to perceptual similarities, cosine similarities are inverted for source-target gender conditions and present male-to-male conversions with the highest similarities (while human perception rates female-to-female as most similar). This inversion between measurements suggests that human listeners may be assigning more importance to higher frequencies over lower frequencies when assessing timbral similarities between voices, giving them a bias towards females due to the increased amount of high-frequency energy. Without analysing the two similarity metrics comparatively, we would have concluded that female related conversions were perceived to be more similar because males have a wider pitch and timbral range due to their ability to move between chest and falsetto voice, and therefore their voices are harder to model. This is a clear example of why having both subjective and objective evaluation methods are vital to assessing models' performances.

From these graphs, we can deduce that the *RECON+BN-LR* condition severely hindered the model's capabilities of producing natural voices in comparison to the *RECON* and *RECON+SIE-LR* conditions. This can be reasoned by the fact that adding a loss component of equal weighting can slow down a network's rate of convergence if it has more objectives to consider optimising for. However it is interesting to note that even though the *RECON+SIE-LR* condition is similar to *RECON+BN-LR* in nature, we see no similar statistically significant drop in performance when compared to *RECON*. This demonstrates that ensuring the decoder prioritises the use of the conditioning SIEs by retaining this information in the output leads towards significantly better voice conversion and naturalness than the use of the bottleneck's non-timbral information.

3.2 Disentanglement

Accuracy results for bottleneck classification layers on the three models are summarised in Table 1 for convenience. The model using only L_{RECON} produced 45% classification accuracy, which shows that while a bottleneck of dimension 256 (16 timesteps by 16 frequency bins) might produce adequate disentanglement for the VCTK speech dataset as reported in [4], this is not the case for the DAMP singing dataset. This means that there is a considerable amount of SIE information still entangled in the bottleneck.

The model using L_{RECON} and L_{BN-LR} achieved the lowest classification accuracy of 23%, which implies that
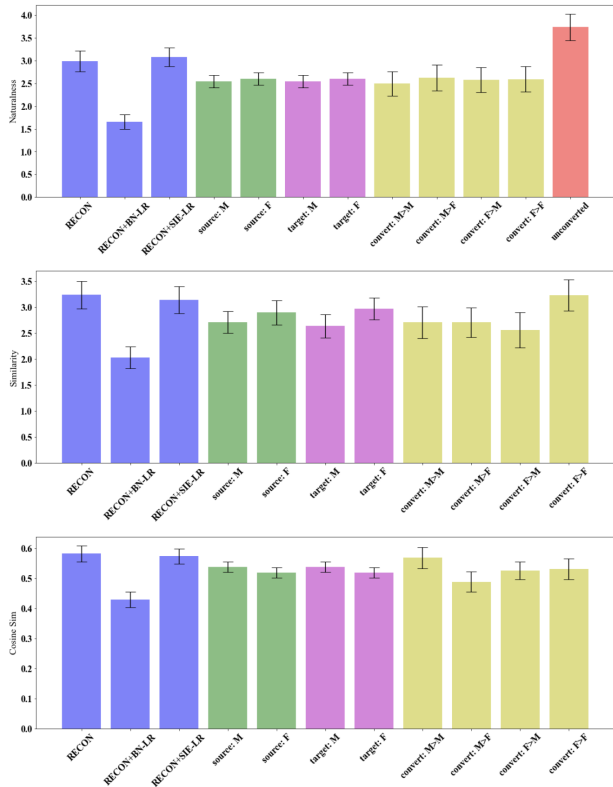
Figure 2. Results for naturalness, similarity and cosine similarity. Blue, green/pink and yellow colours represents condition groups of the model type, source/target genders and source-target gender pairs respectively.

the network's encoder encoded minimal singer identity information and therefore achieves maximal disentanglement between singer identity and non-timbral information. This is likely due to the fact that non-timbral content is prioritised as the information that needs to be stored in the bottleneck due to the conditioning SIEs already providing singer identity information. SIEs may also vary between original and reconstructed representations due to the use of averaged SIEs, which is ignored in the *RECON+BN-LR* condition.

The model using L_{RECON} and L_{SIE-LR} loss produced an accuracy of 35%, which is 10% less than the model trained with only L_{RECON}. This reduction reflects better disentanglement, and is likely because the network is relying more heavily on the decoder utilising the conditioning SIE. It does not however ensure that the encoder avoids encoding singer-identity information, which is why accuracy is still higher than the model using *RECON+BN-LR*. The *RECON+SIE-LR* condition had however been shown to achieve the best singer identity conversion, as seen in Fig. 2, which demonstrates that this loss component offers the added advantage of being robust against poor disentanglement in the bottleneck. This shows that utilising the SIE-LR loss avoids the necessity of manually calibrating the bottleneck's capacity.

Loss components used	Classification accuracy
RECON	45%
RECON and SIE-LR	35%
RECON and BN-LR	23%

Table 1. Classification accuracy results for models using different loss functions.

4. CONCLUSION

We challenged the use of a latent regressor loss on bottleneck embeddings, and confirmed that in an SIE-conditioned autoencoder system, using a latent regressor loss w.r.t. singer identity embeddings rather than bottleneck embeddings led towards voice-converted audio that was more natural and successfully converted, according to both human perception and computational metrics. The advantage of including this loss ensures a more explicit requirement for good conversion, and also requires no calibration of the bottleneck dimensions. While including an SIE latent regressor loss has not been shown to perform better than excluding it, we propose that this precautionary loss may be more advantageous when training an SVC network for longer on larger or more complex datasets which will be addressed in future work. We also proposed the use of a pitch-matching algorithm to ensure evaluations are not misled by diversity in pitch registers between source and target singer recordings (a consideration that is surprisingly absent from previous SVC literature). Objective and subjective ratings for similarity were inverted across gender conditions, which demonstrates the importance of using both methods for evaluations as it is important to understand how a network models data, while considering that the ultimate evaluation relies on human perception. Audio examples and code for this research can be found online [3], where we also present audio clips generated from our *RECON+SIE-LR* model when trained on the entirety of the DAMP dataset for 1 million iterations.

5. REFERENCES

[1] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "FastSVC: Fast Cross-Domain Singing Voice Conversion with Feature-Wise Linear Modulation," in *Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2021, pp. 1–6.

[2] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing Voice Conversion with Disentangled Representations of Singer and Vocal Technique Using Variational Autoencoders," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3277–3281.

---

[3] trebolium.github.io/autosvc_sie-lr_audio

[3] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations," in *Interspeech 2018*. ISCA, 2018, pp. 501–505.

[4] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 2019, pp. 5210–5219.

[5] S. Nercessian, "Zero-shot Singing Voice Conversion," in *Proceedings of ISMIR*, 2020, pp. 70–76.

[6] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2016, pp. 1–6.

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proceedings of Machine Learning Research*, vol. 80, 2018, pp. 4693–4702.

[8] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, "Unsupervised Cross-Domain Singing Voice Conversion," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 801–805.

[9] X. Li, S. Liu, and Y. Shan, "A Hierarchical Speaker Representation Framework for One-shot Singing Voice Conversion," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 4307–4311.

[10] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6284–6288.

[11] Z. Tan, J. Wei, J. Xu, Y. He, and W. Lu, "Zero-Shot Voice Conversion with Adjusted Speaker Embeddings and Simple Acoustic Features," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5964–5968.

[12] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network," in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5279–5283.

[13] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv preprint arXiv:1607.08022*, Nov. 2017.

[14] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, Dec. 2021.

[15] M. Pasini, "MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms," *arXiv preprint*, Oct. 2019.

[16] J. Lu, K. Zhou, B. Sisman, and H. Li, "VAW-GAN for Singing Voice Conversion with Non-parallel Training Data," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2020, pp. 514–519.

[17] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3364–3368.

[18] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv preprint arXiv:1711.11293*, Nov. 2017.

[19] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "WGANSing: A Multi-Voice Singing Voice Synthesizer Based on the Wasserstein-GAN," in *2019 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019, pp. 1–5.

[20] Y. Mroueh, T. Sercu, and V. Goel, "McGan: Mean and Covariance Feature Matching GAN," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 2527–2535.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Advances in Neural Infrmation Processing Systems*, vol. 29, pp. 2234–2242, 2016.

[22] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of the 33rd International Conference on Machine Learning*. PMLR, Jun. 2016, pp. 1558–1566.

[23] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[24] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," *arXiv preprint arXiv:1908.06593*, p. 8, 2019.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 2242–2251.

[26] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96.

[27] M. Amodio and S. Krishnaswamy, "TraVeLGAN: Image-To-Image Translation by Transformation Vector Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 8975–8984.

[28] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning Dense Correspondence via 3D-Guided Cycle Consistency," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 117–126.

[29] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial Feature Learning," in *International Conference on Learning Representations*, 2017.

[30] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing Voice Conversion Network with Cycle Consistency Loss of Speaker Identity," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 507–513.

[31] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 496–500.

[32] B. O'Connor, S. Dixon, and G. Fazekas, "Zero-shot Singing Technique Conversion," in *Proc. of the 15th International Symposium on CMMR*, Tokyo, 2021, pp. 235–244.

[33] D.-Y. Wu and H.-y. Lee, "One-Shot Voice Conversion by Vector Quantization," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7734–7738.

[34] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Content Based Singing Voice Extraction from a Musical Mixture," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 781–785.

[35] S. Wager, G. Tzanetakis, S. Sullivan, C.-i. Wang, J. Shimmin, M. Kim, and P. Cook, "Intonation: A Dataset of Quality Vocal Performances Refined by Spectral Clustering on Pitch Congruence," in *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 476–480.

[36] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4879–4883.

[37] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh, Centre for Speech Technology Research (CSTR), Tech. Rep., 2016.

[38] B. O'Connor, S. Dixon, and G. Fazekas, "An Exploratory Study on Perceptual Spaces of the Singing Voice," in *Proceedings of 2020 Joint AI Conference on Music Creativity*, vol. 1, Stockholm, Sweden, Oct. 2020.