

Music2Fail: Transfer Music to Failed Recorder Style

Chon In Leong, I-Ling Chung, Kin-Fong Chao,
Jun-You Wang, Yi-Hsuan Yang, Jyh-Shing Roger Jang
National Taiwan University, Taiwan

E-mail: r11922174@csie.ntu.edu.tw, r11942103@ntu.edu.tw, r12922179@ntu.edu.tw,
junyou.wang@mirlab.org, yhyangtw@ntu.edu.tw, jang@csie.ntu.edu.tw

Abstract—The goal of music style transfer is to convert a music performance by one instrument into another while keeping the musical contents unchanged. In this paper, we investigate another style transfer scenario called “failed-music style transfer”. Unlike the usual music style transfer where the content remains the same and only the instrumental characteristics are changed, this scenario seeks to transfer the music from the source instrument to the target instrument which is deliberately performed off-pitch. Our work attempts to transfer normally played music into off-pitch recorder music, which we call “failed-style recorder”, and study the results of the conversion. To carry out this work, we have also proposed a dataset of failed-style recorders for this task, called “FR109 Dataset”. Such an experiment explores the music style transfer task in a more expressive setting, as the generated audio should sound like an “off-pitch recorder” while maintaining a certain degree of naturalness.¹

I. INTRODUCTION

Generally, the goal of *music style transfer* is to change the style of the input audio while preserving the content of the input audio². In particular, the *content* of the input music data refers to music features such as rhythm and melody, while the *style* refers to the unique perceived characteristics that an instrument expresses in music performance. According to the wide range of conversion goals and targets, the tasks of music style transfer can be broadly classified into three categories: *one-to-one* [1], [2], *many-to-one* [3], [4], *many-to-many* [5], [6]. Several music style transfer methods have been inspired by research from different fields, such as voice conversion (VC) [5]–[8] and image-based style transfer [9], [10].

Most works mainly use well-played audio as training data for both the source and target domains. This implies that well-played input audio will be converted into equally well-played output audio. Although this ensures faithful style transfer, assuming all audio to be well-played in the real world restricts the expressiveness of instruments. The Singing Voice Beautifying (SVB) task [11] is a special case. They used paired data of amateur and professional singing voices for training, aiming to correct the pitch and improve the vocal tone, indicating that the source audio was not well-played. In this paper, we focus on another special case, where the target domain is not well-played. For example, the target domain may be a soprano

recorder that is deliberately performed poorly³. We refer to this case as *failed-music style transfer*. The motivation is that such failed music contains a wider range of characteristics, but humans can still distinguish between a “failed recorder” and another instrument. This poses a more difficult scenario to music style transfer: Can a music style transfer model not only tackle well-played instruments but also instruments that are *not* well-played?

Take a soprano recorder as an example, a fail-style recorder may contain many types of errors, such as:

- **Cracked voice.** Producing a harsh sound.
- **Weird dynamics.** Unnatural volume while playing.
- **Failed tonguing.** Mistakes in the articulation.
- **Overblowing.** Blowing too hard, causing the voice to sound raspy.
- **Underblowing.** Blowing not hard enough, causing the voice to sound hissing.

Generally, these are considered errors that should not occur in live performances. However, by definition, such errors should not make a style transfer model malfunction. Instead, a style transfer model should generate audio that sounds like a failed recorder (sometimes has an unpleasant style, but still sounds like a recorder). Such failed music style transfer might be useful in the fields of entertainment, it could serve as the score for some comedies or some humorous scenes.

In this paper, we investigate the music style transfer scenario of failed recorders, treating it as a type of instrument. We apply various general style transfer methods and analyze the conversion results. However, there are no existing datasets for such scenarios which were deliberately recorded as failed-style. To facilitate our research, we propose the “FR109” dataset, a collection of failed-style recorder music recorded by a professional, with deliberately included failures.

To sum up, the main contributions of this paper are two-fold:

- We discuss the special scenario of *failed-music style transfer* that serves as a more challenging task for style transfer. Regarding the experimental results, we analyzed them from the perspectives of the Mel spectrogram and Wiener entropy, providing corresponding analyses and interpretations.
- To carry out the work for *failed-music style transfer*, we propose the FR109 dataset, a dataset of failed recorder

¹Our demo website is now available on: <https://navi0105.github.io/demo/music2fail/>

²In practice, due to the differences in pitch ranges of different instruments, we may perform additional pitch shifting in the experiments.

³A famous example can be found at <https://www.youtube.com/watch?v=X2WH8mHJnhM>

performance, created intentionally by an experienced individual playing a recorder.

II. DATASETS

In this work, we adopted three datasets in the experiments, including two publicly available datasets (URMP [12] and Bach10 [13]), and our FR109 dataset. The comparison of different datasets is shown in Table I.

A. The URMP dataset

The URMP dataset [12] contains 44 music pieces ranging from duets to quintets, with separated tracks for individual instrument recordings. There are 14 distinct instruments in this dataset. In our experiments, we only used the violin, clarinet, and saxophone tracks as the training data.

B. The Bach10 dataset

The Bach10 dataset [13] consists of audio recordings of 10 J.S. Bach chorales performed separately with violin, clarinet, saxophone, and bassoon. In our experiments, we only used the violin, clarinet, and saxophone tracks as the testing data. Since the training data (URMP) and testing data (Bach10) belong to different datasets, such an evaluation scenario is more challenging.

C. The proposed FR109 dataset

As for the failed-music style transfer, we proposed the FR109 dataset, which consists of 109 songs recorded with a soprano recorder played by a professional, with a total duration of 5.05 hours. Errors are introduced to each performance intentionally.

As discussed in Section I, the types of errors include cracked voice, weird dynamics, failed tonguing, overblowing, and underblowing. To compute the statistics of the dataset, we extract the pitch of recorder music using CREPE [14]. The pitch mean of the FR109 dataset is around 905Hz (between A5 and A#5), and the maximum pitch value is 1990Hz (around B6). These statistics match the actual pitch range of the soprano recorder, which spans from C5 to D7.

Since there is no other dataset for failed recorders, we use the FR109 dataset as both the training dataset and the testing dataset in the failed-music style transfer experiments. A 90%/10% split is employed to divide the dataset into a training dataset and a testing dataset. In our experiments, we trained style transfer models to perform style transfer between all 4 instruments (violin, clarinet, saxophone, and failed recorder).

We are making the FR109 dataset publicly available for reproducibility, you can check out more information on our official GitHub page: <https://github.com/navi0105/Music2Fail>

III. METHOD

In this work, we experiment with three different well-known style transfer methods for failed-music style transfer, they are StarGAN [15], VAE-GAN [7], and DDSP [3].

StarGAN [15] introduced domain labels to the generator and discriminator, the generator uses the domain label to specify

TABLE I
THE LIST OF DATASETS UTILIZED IN THIS WORK, INCLUDING THE PROPOSED FR109 DATASET. "PIECES" STANDS FOR THE NUMBER OF MUSIC PIECES.

Dataset	Instrument	Pieces	Total duration
URMP	Violin	34	1.02 hours
	Clarinet	10	0.30 hours
	Saxophone	11	0.26 hours
Bach10	Violin	10	0.09 hours
	Clarinet	10	0.09 hours
	Saxophone	10	0.09 hours
FR109	(Failed) recorder	109	5.05 hours

the target domain, while the discriminator needs to predict the input's domain. During training, the generator and the discriminator contest with each other, the generator's objective is to fool the discriminator, making it mispredict the domain label, while the discriminator's objective is to avoid being fooled by the generator. StarGAN only used a single generator and discriminator for learning a multi-mapping between different styles, instead of one generator and one discriminator for each pair of styles.

VAE-GAN [7] used one generator and one discriminator for each domain, the generator used a variational autoencoder composed of two parts: the universal encoder and a decoder, the universal encoder shared across each generator to encode the input to latent code, and a decoder to transfer the latent code to the target domain. Since it uses the same encoder for every input domain and target domain, the performance was increased due to the variation of the input data. The decoder is domain-specific so it can be specialized to that domain. The discriminator only needed to predict whether the data was generated for that domain.

DDSP [3] integrates classic signal processing with deep learning. This method employs an autoencoder architecture for style transfer within a single domain. The encoder extracts key features from the source audio, including loudness, fundamental frequency, and residual information, while the decoder maps these features to control parameters for synthesizers to generate the output audio. DDSP has an assumption that the pitch component extracted from the source audio should closely match the fundamental frequency of the output audio, which may not be suitable for failed-music style transfer.

StarGAN and VAE-GAN are two-stage style transfer pipelines, where we first convert the source audio to Mel spectrogram. Then, the Mel spectrogram was transferred to the failed recorder style using the generator. Finally, the vocoder generates waveform from the transferred Mel spectrogram. Here, we use BigVSAN [16] as our vocoder, its pretrained weight are available in their official repository⁴, which was pretrained on the LibriTTS dataset's training dataset [17] for 10 million steps.

⁴<https://github.com/sony/bigvsan>

TABLE II
THE FAD METRICS OF STYLE TRANSFER
MODELS TO FAILED RECORDER IN THE
BACH10 DATASET.

Models	FAD (\downarrow)
StarGAN	13.87
VAE-GAN	7.27
DDSP	38.91

IV. EXPERIMENTS

As discussed in Section II, we combined the URMP dataset [12] and the FR109 dataset’s training dataset for training and the Bach10 dataset [13] for evaluation. Three different methods are compared in the experiments, they are StarGAN [15], VAE-GAN [7], and DDSP [3].

A. Training

1) *Data preprocessing*: We refer to the arguments for calculating the Mel spectrogram from BigVSAN [16] to compute the Mel spectrograms of music data in our dataset, 24,000 for sampling rate, 100-bands of Mel filter bank, 1024 for FFT / Hann window, hop size is 256 and the frequency range is from 0 to 12,000 Hz.

B. Evaluation

In this section, we compare the performance between StarGAN, VAE-GAN, and DDSP. Both objective and subjective experiments are conducted.

1) *Objective evaluation*: Fréchet Audio Distance (FAD) [18] is a reference-free metric to compute the Fréchet Inception Distance (FID) between audio embedding sets extracted from the reference set and evaluation set. In the experiments, the reference set is music from the testing dataset and the evaluation set is music generated by the model. FAD represents the degree of dissimilarity between the two sets. The audio embeddings are extracted by a pretrained VGGish audio classification model [19]. We use FAD as an objective evaluation metric to assess the distance between the audio files converted by StarGAN / VAE-GAN / DDSP and the real audio performance of a target instrument.

Table II shows the FAD score of Bach10 dataset’s music converted to failed-style recorder music using these three different models. The results indicate that StarGAN performs slightly worse than VAE-GAN on both datasets. Considering that StarGAN only utilizes one unified decoder while VAE-GAN uses one decoder for each instrument, such a performance gap is acceptable.

As for the DDSP model, results show that DDSP has a significantly larger FAD compared to StarGAN and VAE-GAN. By inspecting the audio converted by DDSP, we found that they contain a large amount of noise, which is the reason for the high FAD. The results of DDSP demonstrate that its assumptions about pitch invariance can lead to better performance on well-played instrument transitions, but do not apply well to our task.

TABLE III
THE MOS OF THE LISTENING TEST ON THE BACH10
DATASET. THE NUMBERS INSIDE THE CELLS REPRESENT
THE MOS AND THEIR STANDARD DEVIATIONS. SS, MS,
AND SQ INDICATE THE STYLE SIMILARITY, MELODY
SIMILARITY, AND SOUND QUALITY, RESPECTIVELY.

Models	SS (\uparrow)	MS (\uparrow)	SQ (\uparrow)
StarGAN	2.54 \pm 1.26	3.15 \pm 1.19	2.46 \pm 1.27
VAE-GAN	2.98 \pm 1.23	3.56 \pm 0.93	3.00 \pm 0.98
DDSP	1.33 \pm 0.77	2.19 \pm 1.11	1.38 \pm 0.70

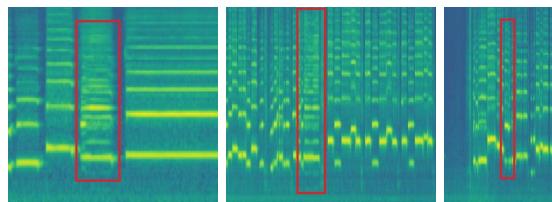


Fig. 1. Mel spectrograms of failed recorder music, the red rectangular parts show the inharmonic partials.

2) *Subjective evaluation*: We performed a listening test that evaluates the performance of converting these three (well-played) instruments into failed recorder in the Bach10 dataset (3 source-target pairs).

For each source-target pair, we randomly choose one audio clip for the listening test. We employed a rating scheme based on the Mean Opinion Score (MOS) [20]. For each audio clip (converted by one of the models), we asked the participants to evaluate its quality in three aspects: (1) *Style similarity* (SS) to the target instrument, (2) *Melody similarity* (MS) to the source audio, (3) *Sound quality* (SQ) of the converted audio. The scoring ranges from 1 to 5, where 1 is the worst and 5 is the best. In total, we received 16 valid responses from the listening test. Table III shows the MOS of the Bach10 dataset.

The results indicate that StarGAN’s overall performance falls behind VAE-GAN’s on all metrics in the conversion to failed recorder. The p -values between StarGAN and VAE-GAN are 0.09 (SS), 0.06 (MS), and 0.02 (SQ). Although only the sound quality (SQ) is considered statistically significant, overall, we can still conclude that StarGAN is slightly inferior to VAE-GAN in converting to failed recorder music. As for DDSP, similar to the objective results, the MOS results of DDSP are significantly worse than those of StarGAN. This is likely because DDSP generates noise more frequently. All the t -test yielded p -values well below 0.05. This reflects the specific challenges involved in music style transfer to failed instrument music for DSP-based synthesizers.

V. ANALYSIS

In this section, we analyze the results of failed recorder style transfer, and further compare the tasks between the conversion to well-played instruments (violin, clarinet, saxophone) and failed recorder.

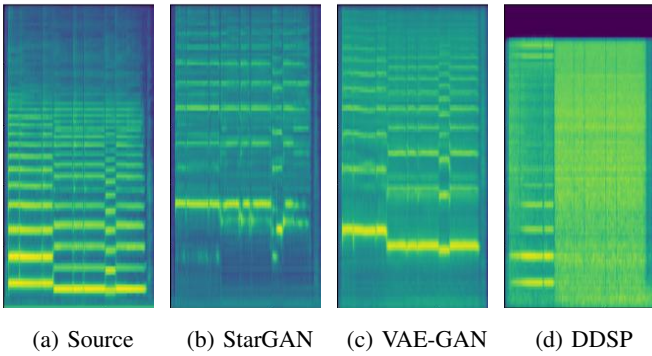


Fig. 2. The Mel spectrograms of a failed-music style transfer example. (a) The Mel spectrogram of the source audio, which is performed by a saxophone; (b) The Mel spectrogram of the converted audio (to failed recorder) by StarGAN; (c) The Mel spectrogram of the converted audio (to failed recorder) by VAE-GAN. (d) The Mel spectrogram of the converted audio (to failed recorder) by DDSP.

TABLE IV
WIENER ENTROPY OF THE URMP DATASET AND FR109 DATASET. VN., CL., SAX., AND REC. REPRESENT VIOLIN, CLARINET, SAXOPHONE, AND RECORDER, RESPECTIVELY.

Dataset	Instrument	Wiener entropy
URMP	Vn. / Cl. / Sax.	0.0005
FR109	Rec.	0.0345

A. Mel spectrogram analysis

To understand the challenge of failed-music style transfer, we first visualize the spectrograms of the failed recorder music in the FR109 dataset, as shown in Figure 1. The red rectangular parts of the Mel spectrograms implies that the sound has *inharmonic partials*, meaning that the frequencies of the overtones do not align with integer multiples of the fundamental frequency. This creates a more complex and less predictable timbre. These inharmonic partials are considered features of the failed recorder because they are present in the Mel spectrograms of every failed recorder sample. Such inharmonic partials are rarely found in well-played instrument performances.

Next, we visualised an example of failed-music style transfer, Figure 2(a) shows the Mel spectrogram of source audio performed by a saxophone, in which there is no clear inharmonic partial. Figure 2(b) and Figure 2(c) show the audio converted to a failed recorder by StarGAN and VAE-GAN, respectively. We can clearly see that inharmonic partials occur throughout the whole Mel spectrogram of StarGAN, showing that it does capture the characteristic of failed recorders and performs style transfer accordingly. For VAE-GAN, inharmonic partials can still be seen, but not as clearly as that of StarGAN. This shows that in this particular case, while both StarGAN and VAE-GAN do perform style transfer to some extent, StarGAN achieves a better style similarity to a failed recorder. Our informal listening test also confirms this observation.

Based on Figure 1 and Figure 2, it can be seen that failed-

TABLE V
WIENER ENTROPY OF THE STYLE TRANSFER RESULTS OF STARGAN, VAE-GAN AND DDSP ON DIFFERENT TARGET INSTRUMENTS. VN., CL., SAX., AND REC. REPRESENT VIOLIN, CLARINET, SAXOPHONE, AND RECORDER, RESPECTIVELY.

Model/Target	Vn.	Cl.	Sax.	Rec.
StarGAN	0.0007	0.0004	0.0002	0.0153
VAE-GAN	0.0003	0.0003	0.0006	0.0159
DDSP	0.0008	0.0008	0.0005	0.2041

music style transfer does show a clearly different characteristic to the style transfer of other well-played instruments. To achieve style transfer to failed music, a model has to generate audio with unique properties that do not usually occur in well-played music. Discussing such a task would help understand the performance and the limitation of a style transfer model in another aspect.

B. Wiener entropy

Furthermore, we utilized the STFT-based *Wiener entropy* [21] to quantify how much the noise-like sound is in the results produced by StarGAN and VAE-GAN, along with the Wiener entropy of the URMP dataset and FR109 dataset, which serve as the benchmark for real performance of well-played music and failed music. Table IV shows the Wiener entropy of the URMP dataset and the FR109 dataset, i.e. well-played instrument music and failed recorder music, we can see that failed recorder music exhibits a higher proportion of noise-like characteristics compared to well-played instrument music. Table V shows the Wiener entropy of each of the models on different target instruments. We can see that when the target instruments are well-played instruments, the noise in the results from StarGAN and VAE-GAN are similar to the URMP dataset since their Wiener entropy is very similar. For failed recorder music, we can see that there is a gap between the Wiener entropy of FR109 and the Wiener entropy of StarGAN or VAE-GAN for converting to failed recorder, this shows that there is still room for improvement in converting music to the failed recorder style.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have conducted a series of experiments on the failed-music style transfer, and analysed the characteristics of this relatively special transfer in different aspects through various evaluations.

Furthermore, we have released the FR109 dataset, consisting of failed recorder performances, which is useful for investigating the expressiveness of different style transfer model. Through this study, we hope to propose a music style transfer task that is different from the usual music style transfer task that pursues sound quality and accuracy, but rather a music style transfer task that is more versatile.

REFERENCES

- [1] M. Pasini, “Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms,” *arXiv preprint arXiv:1910.03713*, Oct. 2019.
- [2] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, “Timbretron: A wavenet(cycleGAN(cqt(audio))) pipeline for musical timbre transfer,” *CoRR*, vol. abs/1811.09620, Nov. 2018.
- [3] J. H. Engel, L. Hantrakul, C. Gu, and A. Roberts, “DDSP: differentiable digital signal processing,” *CoRR*, vol. abs/2001.04643, Jan. 2020.
- [4] S. Nercessian, “Differentiable WORLD synthesizer-based neural vocoder with application to end-to-end audio style transfer,” *CoRR*, vol. abs/2208.07282, 2022. [arXiv: 2208.07282](https://arxiv.org/abs/2208.07282).
- [5] A. Bitton, P. Esling, and A. Chemla-Romeu-Santos, “Modulated variational auto-encoders for many-to-many musical timbre transfer,” *CoRR*, vol. abs/1810.00222, Oct. 2018.
- [6] Y. Wu, Y. He, X. Liu, Y. Wang, and R. B. Dannenberg, “Transplayer: Timbre style transfer with flexible timbre control,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.
- [7] R. S. Bonnici, M. Benning, and C. Saitis, “Timbre transfer with variational auto encoding and cycle-consistent adversarial networks,” in *2022 International Joint Conference on Neural Networks*, Padua, Italy, 2022, pp. 1–8.
- [8] L. Comanducci, F. Antonacci, and A. Sarti, “Timbre transfer using image-to-image denoising diffusion models,” *arXiv preprint arXiv:2307.04586*, Jul. 2023.
- [9] T. Kaneko and H. Kameoka, “CycleGAN-vc: Non-parallel voice conversion using cycle-consistent adversarial networks,” in *2018 26th European Signal Processing Conference*, 2018, pp. 2100–2104.
- [10] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [11] J. Liu, C. Li, Y. Ren, Z. Zhu, and Z. Zhao, “Learning the beauty in songs: Neural singing voice beautifier,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, 2022, pp. 7970–7983.
- [12] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,” *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2018.
- [13] Z. Duan, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” *IEEE Trans. Speech Audio Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.
- [14] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.
- [15] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, Utah, USA, 2018, pp. 8789–8797.
- [16] T. Shibuya, Y. Takida, and Y. Mitsufuji, “Bigvsan: Enhancing gan-based neural vocoders with slicing adversarial network,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, 2024, pp. 10 121–10 125.
- [17] H. Zen, V. Dang, R. Clark, *et al.*, “Libritts: A corpus derived from librispeech for text-to-speech,” in *The 20th Annual Conference of the International Speech Communication Association*, Graz, Austria: ISCA, 2019, pp. 1526–1530.
- [18] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A metric for evaluating music enhancement algorithms,” *arXiv preprint arXiv:1812.08466*, Dec. 2018.
- [19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, *et al.*, “Cnn architectures for large-scale audio classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, Louisiana, USA, 2017, pp. 131–135.
- [20] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives,” *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.
- [21] J. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.