Timbre Transfer of Monophonic Stringed Bass Using Variational Autoencoder-Generative Adversarial Network

Trisha Mae N. Cua
Digital Signal Processing Laboratory
University of the Philippines Diliman
Quezon City, Philippines
trisha.mae.cua@eee.upd.edu.ph

Yani P. Sta. Maria
Digital Signal Processing Laboratory
University of the Philippines Diliman
Quezon City, Philippines
yani.sta.maria@eee.upd.edu.ph

Daniel Nikko M. Tumanut
Digital Signal Processing Laboratory
University of the Philippines Diliman
Quezon City, Philippines
daniel.nikko.tumanut@eee.upd.edu.ph

Abstract—**This research project investigated three designs of Variational Autoencoder-Generative Adversarial Network (VAE-GAN) to perform musical timbre transfer of monophonic samples of stringed bass instruments. Music samples from the bass instrument class for training the model were obtained from the NSynth dataset. Audio features are represented using Constant-Q Transform (CQT) spectrograms and Mel-spectrograms. An image-based style transfer using VAE-GAN, with respect to the timbre of the vocoded target audio, was applied to the CQT spectrograms and mel-spectrograms then the output spectrogram with a new timbre was converted back to audio waveform using 1) a conditional Griffin-Lim algorithm and 2) a MelGAN vocoder. Finally, two evaluation methods were used: 1) Dynamic Time Warping (DTW) for the objective test and 2) ABX subjective listening test. Objective testing results show that the architecture of Melspec-MelGAN has the smallest average dynamic time warping distance with a DTW score of 1814.37 units, with the output waveform exhibiting distortions in the generated waveform. Subjective testing results show that the architecture of Melspec-Griffin Lim is the most viable method for timbre transfer based on the respondent's choice of approximately 80.8% and it was also reflected in the analysis of the spectrograms.**

Index Terms—**Timbre Transfer, VAE-GAN, CQT, Mel-spectrogram, MelGAN, Griffin-Lim, DTW, Monophonic stringed Bass instruments**

I. INTRODUCTION

Timbre transfer deals with modifying audio samples such that the underlying timbre is altered while its content is preserved. This allows a source instrument to be manipulated such that it may sound like a target instrument has played an audio sample originally played by that source instrument. Application of successful timbre transfer may prove to be beneficial in areas such as music production, music enhancement, voice anonymization, and data augmentation. Before the modification in timbre transfer can be achieved, the challenge lies in how timbre can be first extracted. The nature of timbre is complex as it is described as “the resonance by which the ear recognizes and identifies a voiced speech sound” or "terminology wastebasket" [1][2]. However, it is officially defined by American National Standards Institute to be the attribute of an auditory sensation that allows it to be distinguished from other sounds having the same pitch and loudness [3].

Although it is considered to be strongly characterized by the frequency spectrum, timbre is still highly dependent on sound pressure and the temporal characteristics of a sound, which implies that timbre is a multidimensional quality [4]. Deep learning (DL) architectures have already been used for feature extraction on a given data and from there execute accurate classification tasks by utilizing a multi-layer artificial neural network (ANN). Such utility of deep learning is helpful in extracting the timbre, as it can theoretically learn multidimensional quality of sound. Timbre transfer on the other hand is a task concerned with modifying audio samples such that the timbre is changed while their semantic content, together with the pitch and loudness is retained.

II. RELATED WORKS

A. Timbre Transfer

Several works on audio style transfer follow the principle of image style transfer models, techniques that are extended on timbre transfer. Works regarding timbre transfer that have represented audio waveforms as images using a time-frequency approach involves the work of Huang et al. [5] which analyzed instrument recordings using constant-Q transform (CQT) that are then visualized through rainbowgrams, using color to encode time derivatives of phase. Similarly, the study of Roche et al. [6] represented audio using short-time Fourier transform (STFT) spectrograms which provide time-localized frequency information as frequency varies over time. Bonnici et al. [7] represented audio using spectrograms converted to mel scale (mel-spectrograms) such that pitches sounded equally distant. As these studies followed an image-based time-frequency procedure, audio is handled indirectly which allows for a less complex processing.

In recent studies, image style transfer is done using Generative Adversarial Networks (GAN), which allow fully unsupervised learning and requires unlabeled data in training and has faster processing speeds [8] but are challenging to train as unsupervised learning is computationally complex, prone to lack full support over the data [9]. As such, other works on timbre transfer [6][7][9] opted for Variational Autoencoders (VAE) to allow semi-supervised learning which is a generally simpler method computational-wise since it makes use of an encoder and a decoder that would map the data and makes use of a decoder to attempt reconstruction of data, thus allowing a semi-supervised learning but less complex algorithm [10].

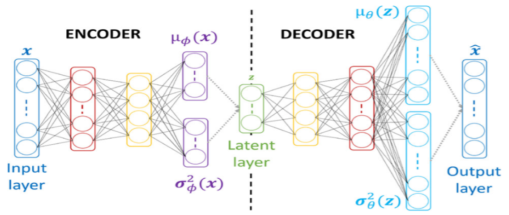


Figure 1.1 — Roche et al. [6] Variational Autoencoder (VAE) architecture

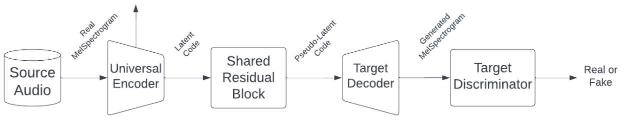


Figure 1.2 — Bonnici [7] Variational Autoencoder-Generative Adversarial Network (VAE-GAN) model

While GANs may offer more overall stability in training, as seen with TimbreTron [5], VAEs are a common and viable method for timbre transfer. For example, Roche et. al. [6] uses the technique of perceptually regularized VAE algorithm with an unsupervised pre-training phase and a fully supervised fine-tuning phase. The VAE model of Roche et.al. (Fig.1.1) makes use of a parametric model for data distribution at the input layer and making use of a standard Gaussian distribution in the vector layers (colored red to purple) and a decoder-network is also implemented using a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ [6]. Another example is the VAE-GAN model of Bonnici et al. [7] (Fig.1.2), similar to a standard VAE algorithm as it only excludes the discriminator at the target domain, which makes use of a one-to-one style transfer [7] with the universal encoder extracts the source audio without consideration to timbre and the target decoder introducing timbre to the target audio, all done in mel-spectrogram format.

B. *Audio Feature Representations*

Audio signals can be represented in the time and frequency domain, enabling the measurement and manipulation of signal properties that vary over time.

1) Mel-Spectrogram: Similar to Fourier transform representations, mel-spectrograms characterize the spectral compositions of a signal over a period of time. This form of representation is closer to human perception of sound as it relates Hertz (f) to the mel (m) scale using the following formula [11].

$$m = 2595 log_{10}(1 + \frac{f}{700})$$

The computation needed for mel-spectrograms is less complex than those of Fourier transform variations, making it ideal for signals that deal with human hearing perception models. However, the process of reconstructing the original signal back from the mel-spectrogram can be lossy as there is missing information between frequencies. Using this representation is beneficial for models that deal with a high number of sound classifications and Architectures that exploit image-based representation of audio signals for processing [12].

2) Constant-Q Transform (CQT): An alternative technique for time-frequency analysis is the Constant-Q Transform (CQT) which uses a logarithmic rather than a linear frequency representation [16]. The analysis introduces a constant frequency resolution ratio, known as the "quality (Q) factor" to create geometrically spaced frequency values. The values are spaced with respect to a pattern $\omega_k = 2^{\frac{k}{b}}\omega_0$ where $k\epsilon\{1, 2, 3, ..k_{max}\}$ and b is a geometric separation constant. The bandwidth of the k_{th} filter is selected as $k = k + 1 - k$.

$$Q = \frac{\omega_k}{\Delta_k} = (2\frac{1}{b} - 1)^{-1}$$

Compared to other methods such as STFT and mel-spectrogram, CQT has a higher spectral resolution even at lower frequencies and is highly effective for convolutional architectures because of approximate pitch equivariance [5].

C. *Waveform Reconstruction*

Works done on timbre transfer usually involve two deep learning models. The first model is concerned with the style transfer on the single or multidimensional representations of audio, and the second model deals with the decoding of the representations back to its audio waveform.

1) Griffin-Lim Algorithm (GLA): The use of the Griffin-Lim algorithm as an audio synthesizer takes advantage of the redundancy present in short-time Fourier transforms in phase reconstruction of signals. GLA does not have prior knowledge of the target signal but relies on the consistency of the spectrogram by iterating two projections throughout a spectrogram. Through multiple iterations, the algorithm recovers complex-valued spectrograms until it converges to a certain spectrogram. The acceleration of said algorithm is presented by Perraudin et al. [14] in what they call the Fast Griffin-Lim algorithm (FGLA) wherein they utilize the difference between the two projections such that the process of obtaining consistent STFT coefficients is optimized. FGLA results in a more accurate and faster convergence but the drawback is losing all the theoretical guarantee of convergence.

2) MelGAN: A non-autoregressive and feed-forward convolutional architecture that generates audio waveforms using GAN patterns. It takes in a spectrogram and produces a raw waveform, making it a lightweight and convenient option for parallelizable audio generation. While WaveNet and WaveGlow produce clearer outputs, MelGAN's flexibility allows for high-quality waveforms with architectural modifications and training. [15][16].

D. *Metrics for Timbre Transfer*

Among the reviewed works, Bonnici et al. [7] used Structural Similarity Index Metric (SSIM) and Frechét Audio Distance (FAD) to assess timbre transfer. A simpler approach that matches signals is a method called Dynamic Time Warping (DTW) which measures the similarity between two signals that may vary in phase or time duration. DTW works by warping the time axis of one signal with respect to the other so optimal alignment can be found. This is useful in cases where the recordings may have slight variations in pitch, tempo, duration, and when distortion and noise are present [17].

Cifka et al. [18] evaluated using the criteria content preservation and style fit. The content preservation part is meant to assess how much of the pitch content of the input is retained in the output, while the style fit criteria is for assessing how well the output fits the target timbre. Following this, Jain et al. [19] utilized a Convolutional Neural Network (CNN) to identify the instruments that generated the spectrograms and was used for predicting the class labels for the generated output images which was compared to the learned classes.

III. Problem Statement and Objectives

A. Problem Statement

Implementations of image-based audio classification paved the way for the exploration of musical timbre transfer. Among the recent timbre transfer algorithms, VAE-GANs were shown to provide a more accurate model through log-likelihood and sampling paired with generative learning characteristics. However, the efficacy of VAE-GANs with different audio feature representations such as Mel-Spectrograms and Constant-Q Transforms as well as their corresponding reconstruction algorithms particularly Griffin-Lim and MelGAN have not yet been explored. This project investigates the combinations of these models to determine and evaluate their performance for the timbre transfer of monophonic stringed bass instruments.

B. Objectives

The objective of this research project is to create a successful timbre transfer system through the following: 1) implement one specific family of instruments (monophonic stringed bass instruments) as a standard basis for timbre transfer, 2) implement Mel-Spectrogram and CQT feature representations, 3) implement a VAE-GAN algorithm, 4) implement Griffin-Lim and MelGAN decoders, and 5) evaluate VAE-GAN algorithm output with ABX testing and Dynamic Time Warping (DTW)

C. Scope and Limitations

This research project focuses on stringed bass instruments as it offers more controllable qualities than instruments with a wide range of pitches. Specifically, the following stringed bass instruments will be used: (1) Electric jazz bass (J-Bass), (2) Electric precision bass (P-Bass), and (3) Double bass, where the double bass is the input or source instrument, while the electric jazz bass and electric precision bass is the target output instruments. With this, the transfer is limited to instruments of the same or close families. Moreover, this kind of setup depicts two one-to-one transfer models as different training is required for each of the three instrument pairs.

IV. Methodology

A. Data Preprocessing

The audio files used in this study were taken from the NSynth dataset which features 4-second monophonic annotated musical notes. Samples from the bass instrument class were obtained to fit the instrument selection of this study which consists of the stringed bass instruments.

The audio inputs were preprocessed as follows: the sampling rate is confirmed to be 16,000 Hz and a root mean square normalization is applied for consistent volume. Mel-spectrograms are obtained with 128 mel frequency bins, 200 hop size, and an 800 Hanning window per frame. CQT spectrograms have 120 frequency bins, 80 hop length, and 24 bins per octave. Spectrograms are logarithmically scaled and min-max normalized, following Bonnici et al. [7], for faster convergence during training. Training, validation, and testing use the same pool of data and are divided into 80%, 10%, and 10%, respectively.

1) Architecture: Utilizing the VAE-GAN model presented by Bonnici et al. [7], which makes use of a one-to-one style transfer with cyclic consistency, three architecture designs are developed for this research project. Each includes the following parts: i) pre-training (preprocessing of inputto match desired audio characteristics), ii) training on VAE-GAN (model learning through respective spectrograms), and iii) analysis of generated output. A total of 500 samples for each instrument is used in this project and this encompasses the data used for training, evaluation, and testing. Only 50 epochs were done in the training phase and was chosen arbitrarily to minimize the chance of overtraining that might cause loss of training data. Varying the audio representation and the vocoder used for each architecture affect the quality of the output and enable the researchers to assess and compare the methods used.

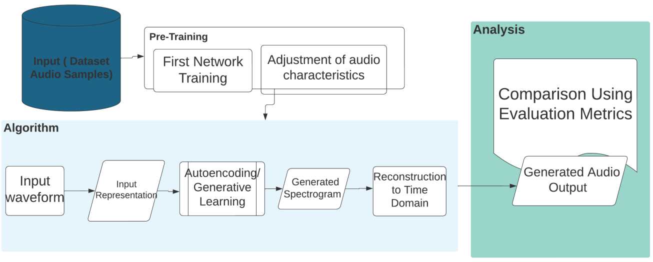

Input (Dataset Audio Samples)
Pre-Training
First Network Training
Adjustment of audio characteristics
Analysis
Comparison Using Evaluation Metrics
Algorithm
Input waveform
Input Representation
Autoencoding/ Generative Learning
Generated Spectrogram
Reconstruction to Time Domain
Generated Audio Output

Figure 2 — VAE-GAN Architecture for Timbre Transfer Flowchart

2) Inference and Vocoding: After the training part comes the inference procedure of the resulting CQT spectrograms and mel-spectrograms with sizes 128x80 and 128x128 respectively. A sliding window with a frame length of 128 traverses the spectrograms with a window overlap of 4 counts per slice. In each slide of the window, the spectrogram is being inferred with the timbre of the target bass instrument. To reconstruct the spectrogram with a new timbre into audio format that can be listened to, the MelGAN vocoder by Kumar et.al [16] and optimized Fast Griffin-Lim algorithm (FGLA) by Perraudin et al. [14] were utilized.

3) Metrics of Evaluation: In evaluating the effectiveness of timbre transfer, both subjective and objective approaches are to be utilized. ABX is a double-blind test methodology that is meant to demonstrate the audible difference between two audio samples. The test has two audio files named “A” and “B” and a respondent will have to discern a third sample X (random version of "A" or "B") if it is more similar to A or B [20]. The statistical significance of ABX testing is based on a binomial distribution with a 95% confidence level to account for errors due to chance.

Alongside this is an evaluation of the output using Dynamic Time Warping which is a method especially useful for comparing pairs of time series data that may vary in length and when time shifts, phase shifts, and distortion are present [21]. The optimal warping path is found using a dynamic programming approach where a cost matrix C(i, j) is computed for each point (i, j), in the source and target signals respectively with, C(0, 0) initialized to D(0, 0) and C(i, 0) and C(0, j) set to infinity. The minimum cost of three possible ways to reach point (i, j) from points (i-1, j), (i, j-1), or (i-1, j-1) is computed using the recurrence relation C(i, j) = D(i, j) + min(C(i-1, j), C(i, j-1), C(i-1, j-1)), and the path with minimum cost is selected.

4) Objective Evaluation Procedure: To objectively validate the results of the testing, a dynamic time-warping (DTW) evaluation code was created through MatLab using the audio toolbox. Synthesized outputs were paired with their respective reference inputs, taken from the dataset, before running through the program. The audio distances for each generated-reference audio pairing were computed and stored in a table and the corresponding mean for the audio distance values were computed.

5) Subjective Evaluation: The subjective testing utilizing the ABX design listening test was accomplished through Google Forms. For a twenty-item questionnaire, 30 participants were asked to choose between two audio files which one better resembles the audio of the target instrument. The audio files consist of 5 random samples of J-Bass and 5 random samples of P-Bass outputs from each of the three architecture designs. The files that are being compared have the same pitch/frequency and have exactly one same processing done to them (i.e., same audio representation used or same vocoder used). In this way, the better sounding audio representation and vocoder is determined by the respondents.

V. Results and Discussion

A. Architecture Outputs

The following are the target visualizations of the architecture and a side-by-side comparison with the input audio, demonstrating the timbre transfer capabilities of each architecture. Both the outputs for J-Bass and P-Bass target instruments are shown in the following figures.

1) CQT & Griffin-Lim: It can be shown in the figures above that the model has been able to retain most of the spectral features of the audio sample more evidently in the lower frequency range while it appears sparse in the higher frequency range for both J-Bass and P-Bass outputs. The audio appears to be full of noise when observed in the time domain, however, it is still evident where exactly in the waveform the bass string was plucked and how it fades out. Moreover, despite the audio sounding to be wobbly, the pitch is retained and heard well which remains consistent with the approximation of pitch equivariance of CQT as described by Huang et al. [5].

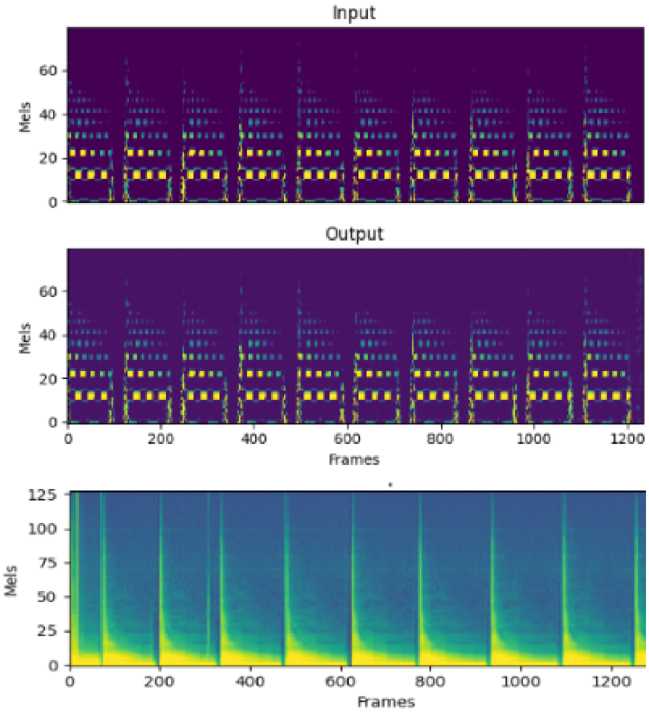


Figure 3.1.—CQT VS Melspec J-Bass Output

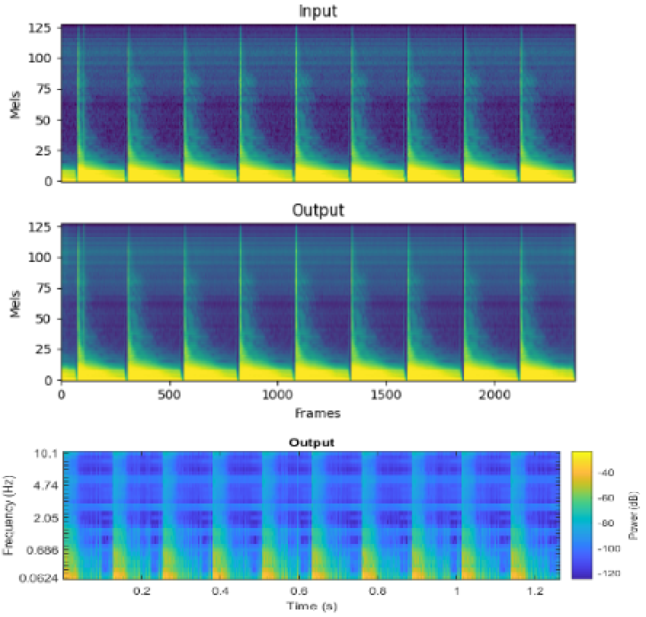


Figure 3.2.—Griffin Lim VS MelGAN P-Bass Output

It can be shown above that for the J-bass output, there is a lower concentration of higher frequencies in the output than the input. The P-Bass output spectrogram, however, retains almost all of the spectral features from the input, which indicates a more accurate reconstruction as the waveform appears smoother and cleaner throughout the whole frequency spectrum. This difference may be attributed to the tonal flexibility of the J-Bass compared to the P-Bass. [22] As there is a greater tonal flexibility in the J-Bass than the P-Bass, this also indicates that there is a wider tonal spectrum that the J-Bass outputs as compared to the P-Bass. From the use of the GLA vocoder, it can be observed that the priority for consistency in obtaining STFT coefficient values allows the waveform to nearly accurately match the pitch of the target output waveform [14].

The mel-spectrograms for the audio synthesized using Mel-

GAN were created in Matlab since the reference architecture did not include options for plotting. It can be seen that both of the audio outputs have been reconstructed according to their reference inputs, particularly at the start of each bin, with degradation as the signal decays. These changes can be seen more in the reconstruction of the P-Bass where the signal decays quicker. The time signal waveform of the J-Bass reconstruction is also more defined than its P-Bass counterpart. With these observations, the MelGAN vocoder displayed its non-autoregressive and convolutional nature that can be refined to become more flexible than the previous, GLA, vocoder.

B. Evaluation

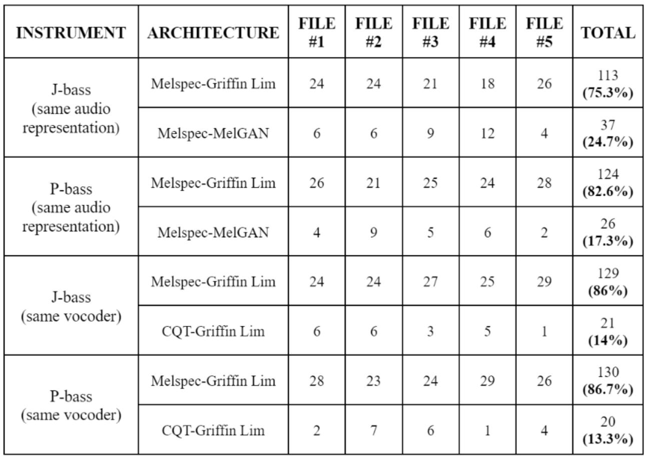
The following are the tables obtained from objective and subjective testing. Shown in the following tables are a summary of the results of the evaluation done, as well as the accompanying analysis of the data obtained.

TARGET INSTRUMENT	DTW MEAN (units)
Melspec-Griffin Lim J-Bass	3415.152193
Melspec-Griffin Lim P-Bass	3005.665443
CQT-Griffin Lim J-Bass	2045.663447
CQT-Griffin Lim P-Bass	2041.960658
Melspec-MelGAN J-Bass	1814.373455
Melspec-MelGAN P-Bass	2466.458281

TABLE I—Objective Evaluation Results

1) Objective Testing: Six pairings from each of the three architectures (with pairs of J-Bass and P-Bass samples) were compared using DTW which returned the average audio distance for each pair of signals (shown above). Among the sets, the Melspec-MelGAN architecture with J-Bass as the target instrument showed the closest audio distance with a calculated mean of 1814.37. At the same time, the Melspec-Griffin Lim architecture also with J-bass as the target instrument showed the least similarity with a calculated mean distance of 3415.15. Ideally, the lower the value of the mean distance, the closer the target instrument is represented. From this observation, the MelGAN vocoder and the Mel-spectral audio representation provided the most viable method of timbre transfer of the dataset among the other methods. However, among the architectures, the Griffin-Lim vocoder output with the same representation sounded the closest to the target instrument as further elaborated in the subjective evaluation.

2) Subjective Testing: A summary of the results from the audio listening test is shown in Table III. They were grouped together by their similarities (same audio representation (mel-spectrogram) or having the same vocoder (Griffin Lim)). Included in the segregation of grouping is the target instrument timbre output, J-Bass or P-Bass.



INSTRUMENT	ARCHITECTURE	FILE #1	FILE #2	FILE #3	FILE #4	FILE #5	TOTAL
J-bass (same audio representation)	Melspec-Griffin Lim	24	24	21	18	26	113 **(75.3%)**
	Melspec-MelGAN	6	6	9	12	4	37 **(24.7%)**
P-bass (same audio representation)	Melspec-Griffin Lim	26	21	25	24	28	124 **(82.6%)**
	Melspec-MelGAN	4	9	5	6	2	26 **(17.3%)**
J-bass (same vocoder)	Melspec-Griffin Lim	24	24	27	25	29	129 **(86%)**
	CQT-Griffin Lim	6	6	3	5	1	21 **(14%)**
P-bass (same vocoder)	Melspec-Griffin Lim	28	23	24	29	26	130 **(86.7%)**
	CQT-Griffin Lim	2	7	6	1	4	20 **(13.3%)**

TABLE II—Subjective Evaluation Results

From the table above, it can be observed CQT-Griffin Lim is not a viable method for timbre transferbas all the results amount to less than 15% of the total votes. And contrary to the objective testing result that Melspec-MelGAN as the viable method for timbre transfer of the dataset, it was Melspec-Griffin Lim that participants have found to be the most viable method for timbre transfer with at least 75% of total votes choosing this method as more viable for timbre transfer than both CQT-Griffin Lim and Melspec-MelGAN. This may be attributed to the performance of the MelGAN vocoder, having more flexibility in attaining the values for recreating the waveform, thus reaching a further range of frequencies matching the ones from the target output. However, the Griffin Lim vocoder maintains the coefficients of the target waveform with STFT, compared to the MelGAN vocoder that maintains the general shape of the waveform but having far less concentration in regards to the fundamental frequencies. In calculating the DTW, it considers the next nearest value in calculating the distances between frequencies of the audio file, regardless of the alignment of the waveforms. The Griffin-Lim vocoder matches the pitch and timbre of the target waveform but has artifacts especially in lower frequencies. The MelGAN vocoder is able to attain some artifacts among all frequencies but loses the concentration of the fundamental frequencies overall and has more distortions in the time domain waveform.

VI. Conclusions

Results on the objective evaluation by computing the DTW mean indicated that the Melspec-MelGAN architecture yielded the lowest mean warping distance of 1814.37 compared to the other architectures. In contrast, results on the subjective evaluation using ABX design listening test on 30 respondents shows that the Melspec-Griffin Lim architecture performed best with a score of 80.8% . The difference in objective and subjective scores may be due to the distortions in the generated output from the Melspec-MelGAN architecture caused by losses in data during training and vocoding.

VII. Recommendations for Future Work

The following are recommendations that can be applied on future work on the topic of timbre transfer on musical instruments: 1) Explore other audio feature representations, 2) Explore other quality metrics (both objective and subjective) for timbre transfer , 3) Increase the number of epochs for training, and 4) Experiment with different types of audio samples, such as musical pieces with pitch variation, to produce more flexible models for testing.

REFERENCES

[1] Merriam-Webster. (n.d.). “Timbre: noun”. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/timbre

[2] Siedenburg, Kai, and Stephen McAdams. “Four Distinctions for the Auditory ‘Wastebasket’ of Timbre1.” Frontiers in Psychology 8 (2017). https://doi.org/10.3389/fpsyg.2017.01747.

[3] “American National Psychoacoustical Terminology.” New York: American National Standards Association, 1973.

[4] Jain, Deepak Kumar, Akshi Kumar, Linqin Cai, Siddharth Singhal, and Vaibhav Kumar. "ATT: Attention-based timbre transfer." In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-6. IEEE, 2020.

[5] Huang, Sicong, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B. Grosse. "Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer." arXiv preprint arXiv:1811.09620 (2018).

[6] Roche, Fanny, Thomas Hueber, Maëva Garnier, Samuel Limier, and Laurent Girin. "Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder." Transactions of the International Society for Music Information Retrieval (TISMIR) 4 (2021): 52-66.

[7] Bonnici, Russell Sammut, Martin Benning, and Charalampos Saitis. "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks." In 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2022.

[8] Ye, Hongliang, and Wanning Zhu. "Music style transfer with vocals based on CycleGAN." In Journal of Physics: Conference Series, vol. 1631, no. 1, p. 012039. IOP Publishing, 2020.

[9] Bitton, Adrien, Philippe Esling, and Axel Chemla-Romeu-Santos. "Modulated variational auto-encoders for many-to-many musical timbre transfer." arXiv preprint arXiv:1810.00222 (2018).

[10] Janetzky, Pascal. “Generative Networks: From Ae to Vae to Gan to Cyclegan.” Medium, July 23, 2021. https://towardsdatascience.com/generative-networks-from-ae-to-vae-to-gan-to-cyclegan-b21ba99ab8d6.

[11] Zhou, Quan, Jianhua Shan, Wenlong Ding, Chengyin Wang, Shi Yuan, Fuchun Sun, Haiyuan Li, and Bin Fang. "Cough recognition based on mel-spectrogram and convolutional neural network." Frontiers in Robotics and AI 8 (2021): 580080.

[12] Thornton, B. Z. J. L. S. "Audio recognition using mel spectrograms and convolution neural networks." (2019).

[13] Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In Proceedings of the IEEE international conference on computer vision, pp. 2223-2232. 2017.

[14] Perraudin, Nathanaël, Peter Balazs, and Peter L. Søndergaard. "A fast Griffin-Lim algorithm." In 2013 IEEE workshop on applications of signal processing to audio and acoustics, pp. 1-4. IEEE, 2013.

[15] Mustafa, Ahmed, Nicola Pia, and Guillaume Fuchs. "Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6034-6038. IEEE, 2021.

[16] Kumar, Kundan, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C. Courville. "Melgan: Generative adversarial networks for conditional waveform synthesis." Advances in neural information processing systems 32 (2019).

[17] Wibawa, I. D. G. Y. A., and I. D. M. B. A. Darmawan. "Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini." In Journal of Physics: Conference Series, vol. 1722, no. 1, p. 012014. IOP Publishing, 2021.

[18] Cífka, Ondřej, Alexey Ozerov, Umut Şimşekli, and Gael Richard. "Self-supervised vq-vae for one-shot music style transfer." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 96-100. IEEE, 2021.

[19] Jain, Deepak Kumar, Akshi Kumar, Linqin Cai, Siddharth Singhal, and Vaibhav Kumar. "ATT: Attention-based timbre transfer." In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-6. IEEE, 2020.

[20] Munson, W. A., and Mark B. Gardner. "Standardizing auditory tests." The Journal of the Acoustical Society of America 22, no. 5_Supplement (1950): 675-675.

[21] Giorgino, Toni. "Computing and visualizing dynamic time warping alignments in R: the dtw package." Journal of statistical Software 31 (2009): 1-24.

[22] Sweetwater. (2022). “Precision VS Jazz Bass: What’s the Difference?”. https://www.sweetwater.com/insync/precision-vs-jazz-bass-whats-the-difference/