ATT: Attention-based Timbre Transfer

Deepak Kumar Jain*, Akshi Kumar†, Linqin Cai§, Siddharth Singhal†‡ and Vaibhav Kumar†‡
*Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing,
College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
§Key Laboratory of Industrial Internet of Things and Networked Control, Ministry of Education,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
†Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India.

***Abstract*—In this paper, we tackle the issue of timbre transfer on a given monophonic music sample. The objective is to change the timbre of source audio from one instrument to another while preserving features such as loudness, pitch, and rhythm. Existing approaches use image-to-image translation techniques on the entire region of time-frequency representations of the raw audio wave, which may lead to the addition of unwanted elements in the final audio waveform. We propose Attention-based Timbre Transfer (ATT), an attention-based pipeline for transferring timbre. To the best of our knowledge, ATT is the first approach which leverages attention for achieving timbre transfer. Further, ATT uses MelGAN for spectrogram inversion, which provides a fast and parallel alternative to other autoregressive music generation approaches, without compromising on the quality. ATT shows promising results, thus efficaciously transferring timbre with minimal offset to other physical characteristics.**

***Index Terms*—Generative Adversarial Networks, Timbre transfer, Style transfer, mel-frequency spectrogram.**

I. Introduction

Music is composed of different elements such as rhythm, dynamics, texture, timbre and melody together creating both short and long-range structure. One of the distinctive features of a musical sound is timbre. The American Standards Association (1960) define timbre as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". They further add, "Timbre depends primarily upon the frequency spectrum, although it also depends upon the sound pressure and the temporal characteristics of the sound", thereby making timbre a multidimensional entity and modelling it a challenging task.

In this paper, we propose a novel attention-based approach for timbre transfer, i.e. given a monophonic music sample, we aim to transfer the timbre of source audio from one instrument to another while preserving features such as loudness, pitch and rhythm. For instance, given a music sample composed using a violin, we aim to obtain an audio waveform for the same composition but with the instrument being a piano. To the best of our knowledge, this is the first application of attention-guided image-to-image translation for transferring timbre.

Further, to convert the translated time-domain representation back to audio, multiple methods can be used. One such method is Griffin-Lim (GL) (Griffin Lim, 1984) which, however, is known to introduce disturbances in the output wave [1]. Recently, neural network-based models such as WaveNet [2] and SampleRNN [3] have been highly successful in generating high-quality audio. However, the inference time of these models is high as the audio needs to be generated sequentially, making them inefficient for real-time usage. Therefore, we have used MelGAN [4], a Generative Adversarial Network based approach for raw audio generation which uses Mel-frequency spectrogram. MelGAN is faster than other Mel-frequency spectrogram inversion methods and can be used for generating audio in parallel. Therefore, making it ideal for real-time applications.

Building upon the above components, we propose attention-based timbre transfer (ATT). For training, ATT uses non-parallel monophonic sound samples. For time-frequency representation, we have used Mel-frequency spectrogram; this is because Mel-frequency distribution is done in a perceptually motivated manner. Empirically, ATT shows promising results, successfully transferring timbre on a given set of monophonic sounds, while preserving other characteristics such as pitch and loudness. Additionally, it achieves high score on a test aimed at classifying musical instrument based upon time-frequency representation.

II. Related works and motivations

A. *Generative Adversarial Network (GAN)*

Generative adversarial network (GAN) [5] is a technique for implicitly learning the distribution of a dataset. The task is formed as a zero-sum game between a generator network(G) and a discriminator network(D). During the training process, G and D are trained sequentially. First, the discriminator learns to minimize the classification loss by assigning a high score to natural data samples and a low score to synthetic data samples. Second, the generator learns to synthesize synthetic data samples such that the score assigned by the discriminator to them increases. The objective function is defined as follows:

$$G^*, D^* = \arg\min_G \max_D -\frac{1}{2}E_{x\sim p_{\text{data}}(x)}[\log D(x)] - \frac{1}{2}E_{z\sim p_z(z)}[\log(1 - D(g(z)))]$$

Where, $p_{\text{data}}(x)$ denotes the real data distribution, $p_z(z)$ denotes a prior distribution such as Gaussian distribution. GANs

‡Authors have contributed equally.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

have been successfully applied to a range of tasks such as image super-resolution [6], image-to-image translation [7], and text-to-image synthesis [8].

B. *Unpaired image-to-image translation*

Unsupervised image-to-image translation techniques aim to create a mapping between the source and the target domains without the presence of paired data samples. In their seminal work, [9] proposed CycleGAN, a generative adversarial network based approach which uses cyclic consistency for learning the desired mapping between two given domains of data. For the source domain X and the target domain Y. The model learns the mappings $F_{Y\to X} : Y \to X$ and $F_{X\to Y} : X \to Y$ along with adversarial discriminators D_X and D_Y, such that D_X distinguishes between the elements of X and translated samples created by $F_{Y\to X}$ and D_Y distinguishes between the elements of Y and artificial samples created by $F_{X\to Y}$. The objective function is given as :

$$\begin{aligned}\mathcal{L}\left(F_{X\to Y}, F_{Y\to X}, D_X, D_Y\right) = \mathcal{L}^x_{\text{GAN}}\left(F_{X\to Y}, D_Y\right) \\ + \mathcal{L}^y_{\text{GAN}}\left(F_{Y\to X}, D_X\right) \\ + \lambda\mathcal{L}_{\text{cyc}}(F_{X\to Y}, F_{Y\to X})\end{aligned}$$

where,

$$\begin{aligned}\mathcal{L}^x_{\text{GAN}}\left(F_{X\to Y}, D_Y\right) = \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] \\ + \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\left(1 - D_Y(F_{X\to Y}(x)\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(F_{X\to Y}, F_{Y\to X}) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}\left[\|F_{Y\to X}(F_{X\to Y}(x)) - x\|_1\right] \\ + \mathbb{E}_{y\sim p_{\text{data}}\,(y)}\left[\|F_{X\to Y}(F_{Y\to X}(y)) - y\|_1\right]\end{aligned}$$

Building upon CycleGAN, [10] propose an attention-guided image-to-image translation technique. The main advantage of using attention-guided architecture is that mapping mechanism translates only those parts of the source image sample which are the most relevant for the translation operation, minimizing changes to any other elements in the background. They introduce attention networks in generators of CycleGAN architecture. The attention networks are trained in tandem with the generator networks. These attention networks are then used to create attention maps for regions which the discriminator networks 'perceives' to be the most discriminative between the given source and target domains. We delineate upon the architectural details in the next section.

C. *Audio style transfer*

Similar to unpaired image-to-image translation, GAN-based methods have also been applied for unpaired audio-to-audio timbre transfer. [11] proposed Timbretron, a timbre transfer pipeline which uses log amplitude Constant-Q-Transform (CQT) for the time-frequency representation of the audio-waveform. A translation between the source and the target CQT is then performed using CycleGAN [9]. Finally, they use WaveNet [2] for generating the target audio waveform from the CQT representation. Since, Timbretron uses CycleGAN for image-to-image translation, unwanted artifacts might be introduced in the target representation. Further, as WaveNet generates audio waveform in a non-parallel manner, it is unable to scale to real-time environments.

GAN based methods have been further applied to voice transfer applications and singing voice separation [12]. Additionally, [13] propose an encoder-decoder approach for successfully translating music across different instruments and styles.

D. *Audio generation from time-frequency representation*

This process involves the creation of audio waveform from its corresponding time-frequency representation such as the Mel-frequency spectrogram. Various techniques have been proposed for the inversion of spectrogram to audio; we briefly highlight them here. One of the most popular approaches for inverting a short-time Fourier transform is Griffin-Lim (Griffin Lim, 1984) method. Griffin-Lim method iteratively estimates the unknown phases by continuously converting back and forth between the time and frequency domain by using short-time Fourier transform and its inverse, replacing the magnitude of each frequency component by that of the predicted magnitude in each iteration. While simplistic, the approach is known to introduce powerful artificial features in the final output as observed by [1].

Wavenet [2] is a fully convolutional, autoregressive model that is capable of producing life-like speech samples. Further, it can also generate high-quality music samples. However, audio samples are required to be generated in a sequential manner which makes the speed of these methods unsuitable for real-time applications.

Recently, GAN based methods have also been used for audio generation. One such approach is MelGAN [4]. MelGAN is a lightweight and parallel GAN architecture for conditional audio synthesis. MelGAN is capable of producing high-quality audio-waveforms parallelly, therefore, making it suitable for real-time applications. Hence, We have employed MelGAN in our approach for constructing the audio waveform from the translated Mel-frequency spectrogram representation.

III. Proposed method

The aim of audio timbre transfer is to construct a mapping from a source audio domain X to a destination audio domain Y such that properties such as pitch and loudness are preserved, while transforming the timbre of the instrument in X to that of target instrument in Y. To this end, we propose **ATT**: **A**ttention-based **T**imbre **T**ransfer pipeline. In this work, we have only considered monophonic sound samples. We intend to work on polyphonic sound samples in the future.

As shown in figure 1, the working of ATT can be divided into three stages:

1) Convert the raw input audio waveform into its equivalent Mel-frequency spectrogram representation.
2) Considering the Mel-frequency spectrogram representation as an image, we model the timbre transfer as an image-to-image translation problem and perform unsupervised attention-guided operation proposed by [10].

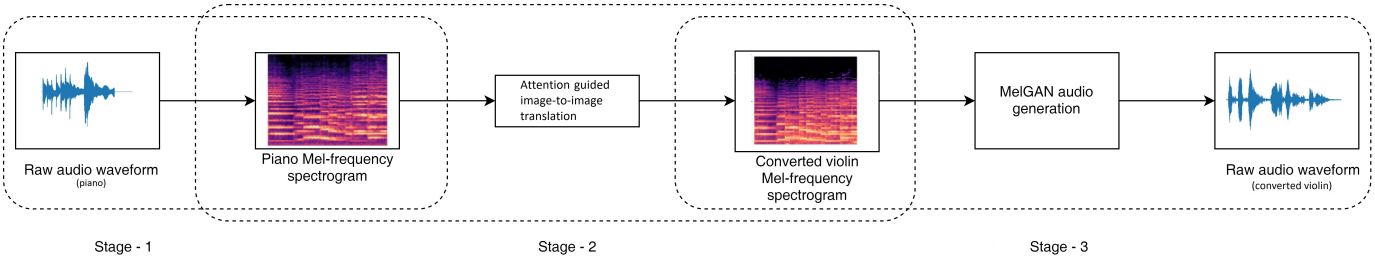

Fig. 1. ATT: Attention-based timbre transfer

3) Finally, the image resulting from the image-to-image translation operation is used as the source in the audio synthesis process using MelGAN [4]. In the subsequent sections, we describe each of these stages in detail.

A. Music representation using Mel-frequency spectrogram

An audio waveform is a variation of pressure with respect to time, carrying the information about changes in amplitude with time. However, utilizing crude audio wave for modelling sound is practically infeasible. This is because crude audio waves are usually sampled at a very high frequency (usually 16Khz or above) for an enhanced temporal resolution, thus, creating a deluge of data, making it unsuitable for modelling purposes. Therefore, modelling methodologies operate upon a simplified representation of the original audio waveform. This representation is lower in resolution than the initial raw wave, but it is easier to work with. In our approach, we have used the Mel-frequency spectrogram for representing the sound. Mel-frequency scale is obtained by the application of non-linear transformation on the linear frequency axis of a spectrogram obtained from short-term Fourier transform (STFT). The transformation results in a frequency scale in which frequencies are represented in a manner which is perceived by the human auditory system, i.e. higher resolution is given to low frequencies and lower resolution is given to high frequencies. Properties such as these have made Mel-frequency spectrogram an attractive choice for training neural networks on large scale audio corpus [14].

B. Timbre transfer on mel-frequency spectrogram Representation

We have employed the attention-guided architecture proposed by [10] for image-to-image translation on the Mel-frequency spectrogram. The advantage of this approach over other methods such as CycleGAN [9] and DualGAN [15] is its ability to discern the most discriminative parts of the image, thereby making changes only to the most relevant parts and adding minimal or no changes to the background. It achieves its objective by creating foreground and background attention maps.

By using attention maps on Mel-frequency spectrograms, we can focus on only those areas of the spectrogram which are most important for discerning the characteristics of the given audio wave. Therefore, When subjected to a image-translation procedure, changes are brought about in only these regions. This reduces the addition of unnecessary features which may deteriorate the quality of audio wave which is obtained from the resulting spectrogram. We now discuss our approach based upon the methodology proposed by [10] in detail.

Having obtained the Mel-frequency spectrogram for the source audio domain X and the target audio domain Y. The attention networks are referred to as A_x and A_y respectively. Each attention network creates attention maps which are denoted as follows: $A_X : X \rightarrow X_a$, $A_Y : Y \rightarrow Y_a$ where X_a, Y_a refer to the attention maps induced by X and Y respectively and consist of [0,1] values per-pixel.

For an input Mel-frequency spectrogram $x \in X$ the transfered representation using the mapping function $F_{X \rightarrow Y}$ and the attention network A_X is defined as :

$$x' = x_a \odot F_{X \rightarrow Y}(x) + (1 - x_a) \odot x \tag{1}$$

where, x_a denotes the attention map $A_X(x)$ and $\odot$ denotes an element-wise product. In the expression above, the former part of the sum denotes the foreground of the image (which is focused upon by the attention map) and the latter part of the sum refers to the background. Figure 2 shows the procedure on actual Mel-Frequency spectrograms. The adversarial loss is then defined as:

$$\mathcal{L}^x_{GAN}\left(F_{X \rightarrow Y}, A_X, D_Y\right) = \mathbb{E}_{y \sim \mathrm{P}_{data}(y)}\left[\log\left(D_Y(y)\right)\right] + \mathbb{E}_{x \sim \mathrm{P}_{data}(x)}\left[\log\left(1 - D_Y\left(x'\right)\right)\right]$$

Further, similar to [9], [10] introduce a cyclic consistency loss which is defined as follows:

$$\mathcal{L}^x_{cyc}\left(x, x''\right) = \left\| x - x'' \right\|_1$$

where, x'' is obtained from x' via $F_{Y \rightarrow X}$ and A_Y in a same manner as equation 1.

The final objective function is then defined as:

$$\mathcal{L}\left(F_{X \rightarrow Y}, F_{Y \rightarrow X}, A_X, A_Y, D_X, D_Y\right) = \mathcal{L}^x_{GAN} + \mathcal{L}^y_{GAN} + \lambda_{cyc}\left(\mathcal{L}^x_{cyc} + \mathcal{L}^y_{cyc}\right) \tag{2}$$

During our training procedure we set the value λ_{cyc} as 10. Further, the input to the discriminator network is also modified using the attention maps so that it consider only attention mapped regions in the synthetic and the real image distribution. Lastly, [10] threshold the learned attention maps in order to prevent attention maps from interfering with the

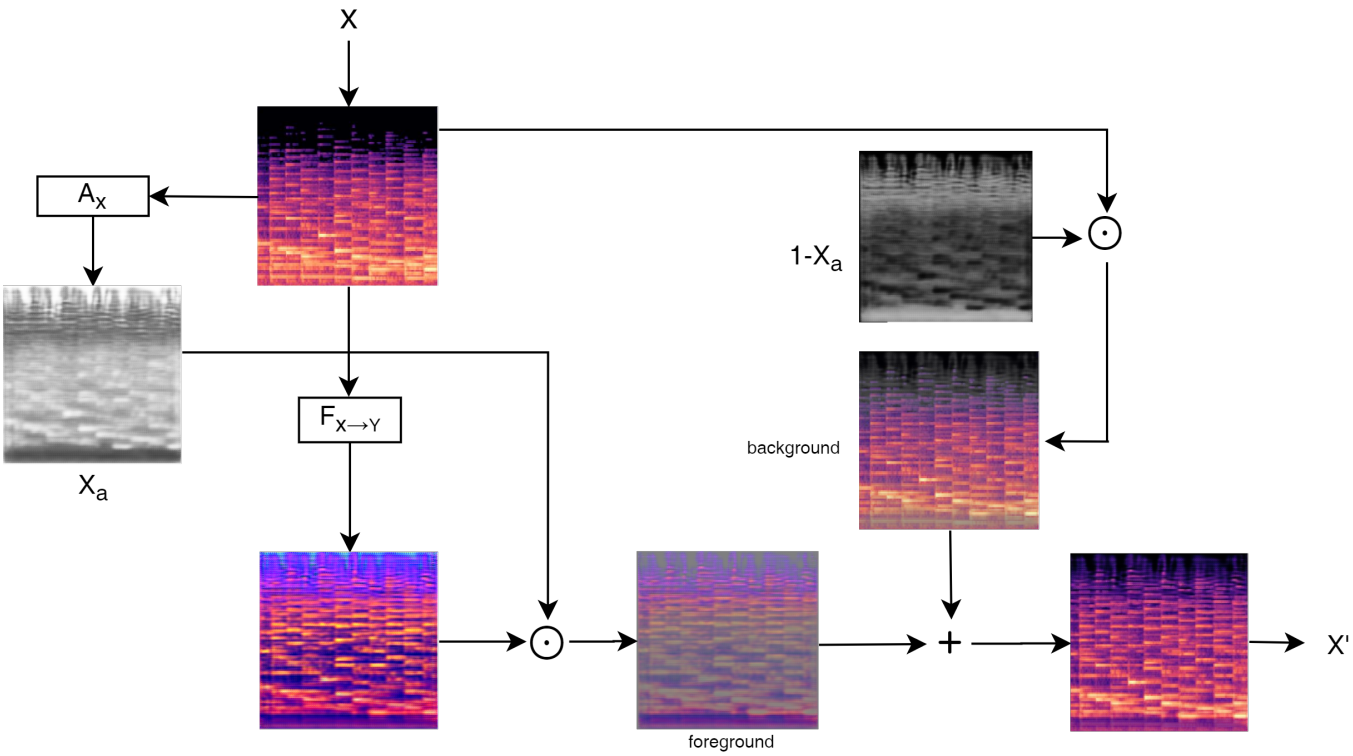

Fig. 2. The generation of output using the mapping function $F_{X\to Y}$ and the attention network A_X. The network works in a symmetric manner for domain Y, mapping function $F_{Y\to X}$, and attention network A_Y.

representation of actual images. We use the threshold value as 0.1 as proposed in the original setup. For further details regarding the architecture, the readers may refer [10].

C. Raw audio waveform generation from mel-frequency spectrogram

After processing the input Mel-frequency spectrogram under the methodology described in the previous section, we obtain the Mel-frequency spectrogram of the audio waveform of the timbre transferred audio. However, the absence of phase information makes it non-trivial to convert the resultant spectrogram into an audio waveform. Therefore, we have used MelGAN for obtaining the audio waveform. MelGAN comprises of generative adversarial networks to produce high-quality audio waveforms given a Mel-frequency spectrogram.

The objective minimized during the training process, as proposed by [4], are defined as follows:

$$\min_{D_k} \mathbb{E}_x \left[\min\left(0, 1 - D_k(x)\right)\right] + \mathbb{E}_{s,z}\left[\min\left(0, 1 + D_k(G(s,z))\right)\right], \forall k = 1,2,3 \quad (3)$$

$$\min_G \left(\mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] + \lambda \sum_{k=1}^{3} \mathcal{L}_{\mathrm{FM}}\left(G, D_k\right) \right) \quad (4)$$

where, G denotes the generator network, D denotes the discriminator network, x is the raw audio waveform, z denotes Gaussian noise vector and s denotes the conditional information (the input Mel-frequency spectrogram) and k denotes the index of the respective discriminator network. $\mathcal{L}_{\mathrm{FM}}$ denotes the feature matching objective, used for reducing the L1 distance of discriminator feature maps of synthetic and real audio. we use $\lambda = 10$ during our training procedure. For further details, the readers may refer [4].

IV. Experiments and Results

A. Datasets

We have used two datasets, namely, MIDI-BACH and real-world dataset for the training and evaluation of our proposed approach.

MIDI-BACH: It is a catalog of J.S Bach's compositions using different Instruments. We use the audio data generated from piano and violin for our analysis.

Real World Dataset: We collected two hours of real world recording from Youtube videos for each instrument [16] [17] [18] [19]. We subsequently converted these video files to audio files. Real World dataset helps us in creating a robust model, effective against noise inherent to real-world audio samples, unlike the MIDI-BACH

These recordings from both the datasets were segmented into six seconds of audio samples. Mel-frequency spectrograms generated for these audio samples were used for training and testing our model. For generating spectrograms, we used 512

samples between successive frames and FFT window of length 2048. The resolution of each spectrogram was [432*432*3]. We generated 2000 Mel-frequency spectrograms for each instrument. We further split this dataset into training, validation, and test set using a 60:20:20 split.

We used an initial learning rate of 0.00005 for the complete network with Adam [20] optimizer for training our network.

B. Evaluation of results

Figures 3 and 6 depict the Mel-frequency spectrograms generated for the audios recorded using piano. These spectrograms were fed as input to our model. Figures 5 and 8 represent the spectrograms for violin generated after attention guided image-to-image translation. [10], using figures 3 and 6 respectively as input. Figures 4 and 7 are the attention maps generated for our first and second input respectively.

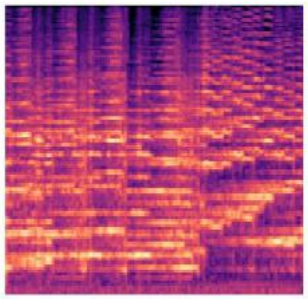

Fig. 3. Input spectrogram of piano: MS_{P1}

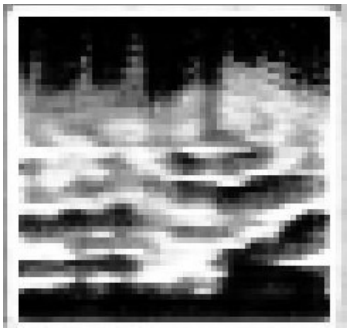

Fig. 4. Generated attention map for MS_{P1}

*1) **Convolutional Neural Network based evaluation**:* We trained a Convolutional Neural Network for identifying the instrument used for generating the input Mel-frequency spectrogram. The classifier was trained on Mel-frequency spectrograms obtained using 1400 audio samples for each instrument category, i.e., piano and violin.

Table I provides a brief overview of the architecture of CNN used during the quality evaluation phase. RGB images [128*128*3] of Mel-frequency Spectrograms were fed as input to the model, and the output was the label for each image.

Our model achieves an accuracy of 99.2% during the testing phase. Figure 9 depicts the Receiver Operating Characteristic

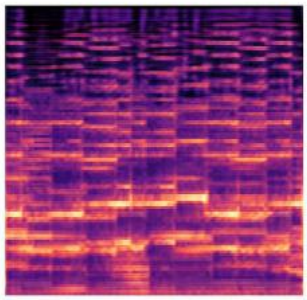

Fig. 5. Generated spectrogram of violin: MS_{P1V}

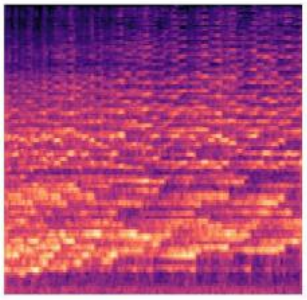

Fig. 6. Input spectrogram of piano: MS_{P2}

curve plot obtained for our Convolutional Network during the testing phase. This represents the high accuracy of our Convolutional Neural Network in identifying the source instrument for the input Mel-frequency Spectrogram.

This CNN was subsequently used for predicting the class labels for our generated images. We obtained an accuracy of 98.89% while predicting the class of our generated spectrograms for a given target label, indicating the good quality and superiority of our proposed architecture in performing timbre transfer.

*2) **MelGAN based evaluation**:* We used MelGAN for reconstructing audio waveform from our generated Mel-frequency spectrograms. These generated audio waveforms were subjected to human listening tests. While the audio-

Conv2d Layers with BatchNorm2d

Filter	Stride	Output	Activation	Max Pooling kernel size
3	1	8	LReLu	2
5	1	32	LReLu	2

Fully Connected Linear Layer

Input Features	Output features
32768	8192
8192	4000
4000	2000
2000	500
500	50
50	2

TABLE I
CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

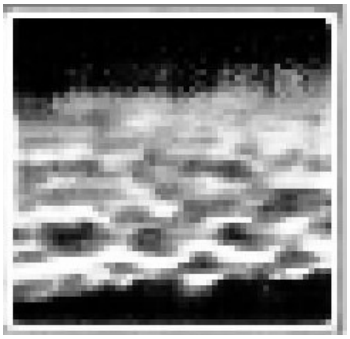

Fig. 7. Generated attention map for MS_{P2}

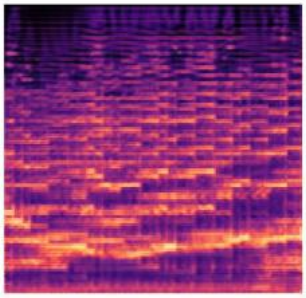

Fig. 8. Generated spectrogram of violin: MS_{P2V}

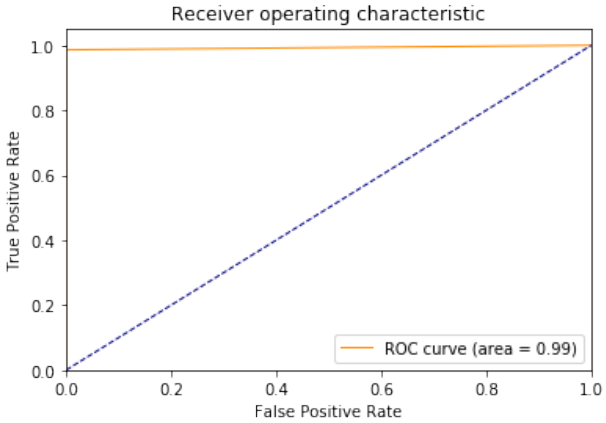

Fig. 9. ROC curve plot for CNN

waveform generated sounded close to its target instrument domain, improvements can be done with regard to reduction of unwanted noises.

V. Conclusion

In this paper, we propose **A**ttention-based **T**imbre **T**ransfer (**ATT**), an attention-based pipeline for musical timbre transfer. ATT is capable of manipulating the timbre of a given monophonic audio sample to that of target instrument while adding a minimal offset to other physical characteristics of sound such as loudness and pitch. By using attention, ATT is capable of manipulating only those regions of the Mel-frequency spectrogram which carry the highest importance for determining the timbre, thereby reducing the addition of unwanted characteristics in the translated image representation. Further, we also use MelGAN [4], which increases the speed of spectrogram inversion process without degrading the quality. In future, we would like to modify the training process for noise reduction as well as extend ATT towards polyphonic sound samples.

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[4] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 881–14 892.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[10] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," in *Advances in Neural Information Processing Systems*, 2018, pp. 3693–3703.

[11] 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," *arXiv preprint arXiv:1811.09620*, 2018.

[12] H. Choi, J.-h. Lee, and K. Lee, ""singing voice separation using generative adversarial networks,"," in *Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017)*, vol. 5.

[13] N. Mor, L. Wolf, A. Polyak, and Y. Taigman, "A universal music translation network," *arXiv preprint arXiv:1805.07848*, 2018.

[14] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[15] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[16] "https://www.youtube.com/watch?v=xkzvya69wco."

[17] "https://www.youtube.com/watch?v=corkefuzsj0."

[18] "https://www.youtube.com/watch?v=0sdlezkik-wt=629s."

[19] "https://www.youtube.com/watch?v=wtbit8alneat=21s."

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.