Instrumental Timbre Transfer Based on Disentangled Representation of Timbre and Pitch

Lin Ye*, Gerald Schuller, Muhammad Imran

Ilmenau University of Technology, Germany
* Now with: Guangdong Smart Home Appliance Innovation Center Co., Ltd., China

Introduction - Background

- Timbre is a perceptual attribute crucial in music synthesis.

  It shapes the unique "color" or quality of sound that allows listeners to distinguish between different instruments playing the same note at the same loudness.

- It is like voice conversion for musical instruments
- Challenges in music synthesis arise due to its complex influencing components.

  Synthesis models must capture the details of timbre like harmonic structures and transient features to replicate an instrument's essence accurately.

Introduction - Our Project Chatbot

- We made a GPT chatbot for the project, which can answer questions quickly
- it speaks any major language

`https://chatgpt.com/g/g-Aag9xbTEk-timbre-transfer-for-musical-instruments`

Introduction - Objective

- This research explores a novel deep learning approach to timbre transfer.
  - Develop a deep learning-based many-to-many music **timbre transfer method**.

    Deep learning-based method can generalize across complex, abstract characteristics and create disentangled, meaningful latent timbre representations.

  - **Disentangle timbre and pitch** for flexible instrumental audio synthesis.

    Disentanglement is key for timbre transfer, as it enables the model to modify timbre while preserving other musical features.

Background - Existing Approaches

- VQ-VAE and continuous-valued embeddings was used to discretely represent content information and style representation in order to create a one-shot timbre transfer model (Cífka et al., 2021).
- Differentiable Digital Signal Processing (DDSP) modules was incorporated into deep learning models to timbre transfer (Engel et al., 2020).
- A many-to-many timbre transfer method was implemented by a pre-trained triplet network composed of a content encoder, a pre-trained style encoder, and a decoder (Chang et al., 2021)

Methodology - Dataset Generation

- Generated using MIDI files and various SoundFont files.

  A MIDI file is a digital file format that contains instructions for creating music, while a SoundFont file contains audio samples and instructions for reproducing sounds of various musical instruments.
- MIDI dataset: Lakh MIDI Dataset (LMD) with hundreds of different timbres (Raffel, 2016).
- Training set: Pairs of 8-second audio samples covering a wide range of music content generated by varying instrument programs.
- Test set: Piano, Guitar, Violin, Saxophone

Methodology - Model Architecture

- Network consists of **content** (pitch) encoder, **style** (timbre) encoder, which are used for corresponding **loss functions**, and decoder.
- Utilizes a beta-Variational Autoencoder (β-VAE) for disentangling timbre and pitch (Higgins et al., 2016).
- Inputs pairs of audio clips with the same timbre but different content.

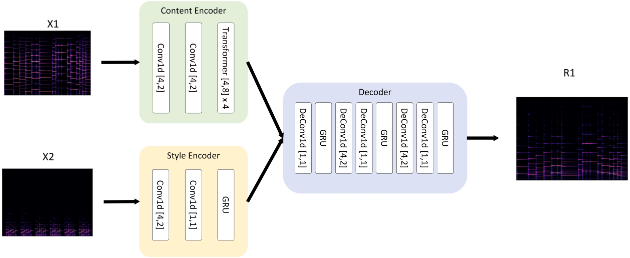


Figure: Model architecture of the timbre transfer network.

Methodology - Loss Function

- Combines conversion error, reconstruction error, and KL divergence.

$$\mathcal{L} = \mathcal{L}_{\text{conv}} + \mathcal{L}_{\text{rec}} + \beta \cdot \mathcal{L}_{\text{kld}}$$

- β parameter controls focus between latent variable independence and reconstruction accuracy.
  - Higher β values emphasize the KL divergence, promoting more disentangled, independent latent features, which helps the model learn distinct and interpretable latent factors.
  - Lower β values reduce the weight of the KL divergence, allowing the model to focus more on accurately reconstructing the input, often at the cost of disentanglement.

Experiments - Audio Results

- Some generated audio examples

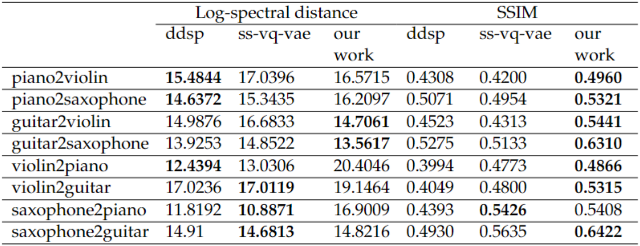
Target timbre	piano	guitar	saxophone	electric guitar
	Play Sound	Play Sound	Play Sound	Play Sound
Input	Play Sound	Play Sound	Play Sound	Play Sound
Generated	Play Sound	Play Sound	Play Sound	Play Sound
Input	Play Sound	Play Sound	Play Sound	Play Sound
Generated	Play Sound	Play Sound	Play Sound	Play Sound

Experiments - Evaluation Metrics

- Spectrogram similarity measured using SSIM (Structural Similarity Index) and log-spectral distance.
- Pitch accuracy calculated using Essentia library's MELODIA algorithm (Raffel et al., 2014).
- FAD score used for overall perceptual quality comparison (Kilgour et al., 2018).
- T-SNE analysis is conducted to assess the model's timbre disentanglement performance (Kilgour et al., 2018).

Results - Timbre Transfer

- Conversion quality assessed through spectrogram comparison.
- Log-spectral distance and SSIM show superior performance of our method.

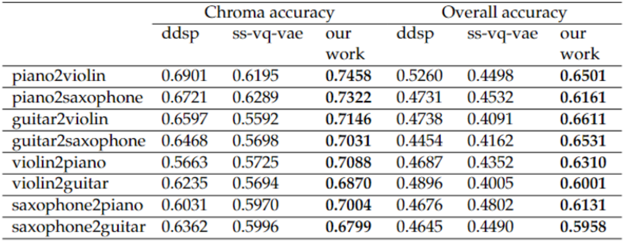


	Log-spectral distance			SSIM		
	ddsp	ss-vq-vae	our work	ddsp	ss-vq-vae	our work
piano2violin	**15.4844**	17.0396	16.5715	0.4308	0.4200	**0.4960**
piano2saxophone	**14.6372**	15.3435	16.2097	0.5071	0.4954	**0.5321**
guitar2violin	14.9876	16.6833	**14.7061**	0.4523	0.4313	**0.5441**
guitar2saxophone	13.9253	14.8522	**13.5617**	0.5275	0.5133	**0.6310**
violin2piano	**12.4394**	13.0306	20.4046	0.3994	0.4773	**0.4866**
violin2guitar	17.0236	**17.0119**	19.1464	0.4049	0.4800	**0.5315**
saxophone2piano	11.8192	**10.8871**	16.9009	0.4393	**0.5426**	0.5408
saxophone2guitar	14.91	**14.6813**	14.8216	0.4930	0.5635	**0.6422**

Figure: Spectrogram comparison between content input and converted audio.

Results - Pitch Accuracy

- Higher pitch accuracy observed for intra-transfer cases.
- Our method demonstrated fewer octave errors than comparison models.



	Chroma accuracy			Overall accuracy		
	ddsp	ss-vq-vae	our work	ddsp	ss-vq-vae	our work
piano2violin	0.6901	0.6195	**0.7458**	0.5260	0.4498	**0.6501**
piano2saxophone	0.6721	0.6289	**0.7322**	0.4731	0.4532	**0.6161**
guitar2violin	0.6597	0.5592	**0.7146**	0.4738	0.4091	**0.6611**
guitar2saxophone	0.6468	0.5698	**0.7031**	0.4454	0.4162	**0.6531**
violin2piano	0.5663	0.5725	**0.7088**	0.4687	0.4352	**0.6310**
violin2guitar	0.6235	0.5694	**0.6870**	0.4896	0.4005	**0.6001**
saxophone2piano	0.6031	0.5970	**0.7004**	0.4676	0.4802	**0.6131**
saxophone2guitar	0.6362	0.5996	**0.6799**	0.4645	0.4490	**0.5958**

Figure: Pitch comparison between content input and converted audio.

Results - Overall Quality (FAD Score)

- FAD score indicates our model generates perceptually closer audio to ground truth.

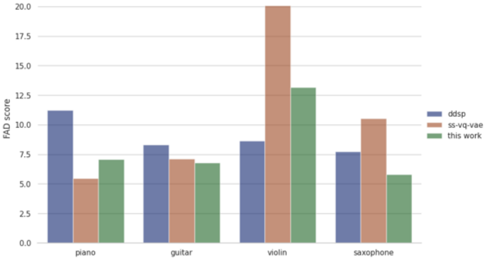
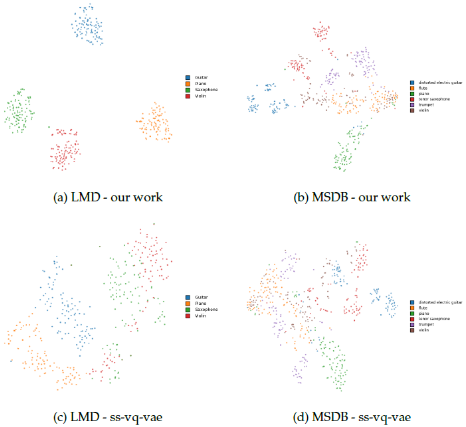


Figure: Perception quality between content input and converted audio.

Results - Disentanglement Performance

- By visualizing the latent style embeddings, we can intuitively see the better disentanglement effect of our model.
- Demonstrates model's capacity for timbre discrimination.



(a) LMD - our work

(b) MSDB - our work

(c) LMD - ss-vq-vae

(d) MSDB - ss-vq-vae

Figure: Perception quality between content input and converted audio.

Discussion

- Limitation: Complex polyphonic structures still pose challenges.

  Polyphonic structures contain multiple, simultaneous harmonics, making it hard for the model to disentangle pitch and timbre accurately for each note independently.

- Model can be improved with 2D convolutional layers and MIDI annotations (Hawthorne et al., 2022).

  The model can better learn the correlations between time and frequency with 2D convolutional layers, while MIDI annotations allow for a clearer representation and manipulation of pitches throughout the audio.

Future Work

- Explore real-time timbre transfer applications.
- Enhance model for broader applications in music synthesis.

References

Chang, Y.-C., Chen, W.-C., and Hu, M.-C. (2021). Semi-supervised many-to-many music timbre transfer. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 442–446.

Cífka, O., Ozerov, A., Şimşekli, U., and Richard, G. (2021). Self-supervised vq-vae for one-shot music style transfer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 96–100. IEEE.

Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (2020). Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., and Engel, J. (2022). Multi-instrument music synthesis with spectrogram diffusion. *arXiv preprint arXiv:2206.05408*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Questions

- Any questions?
- Or scan the QR code to talk with our chatbot on this topic.

`https://chatgpt.com/g/g-Aag9xbTEk-timbre-transfer-for-musical-instruments`
Our project Github repository is:
`https://github.com/TUIlmenauAMS/timbre-transfer`