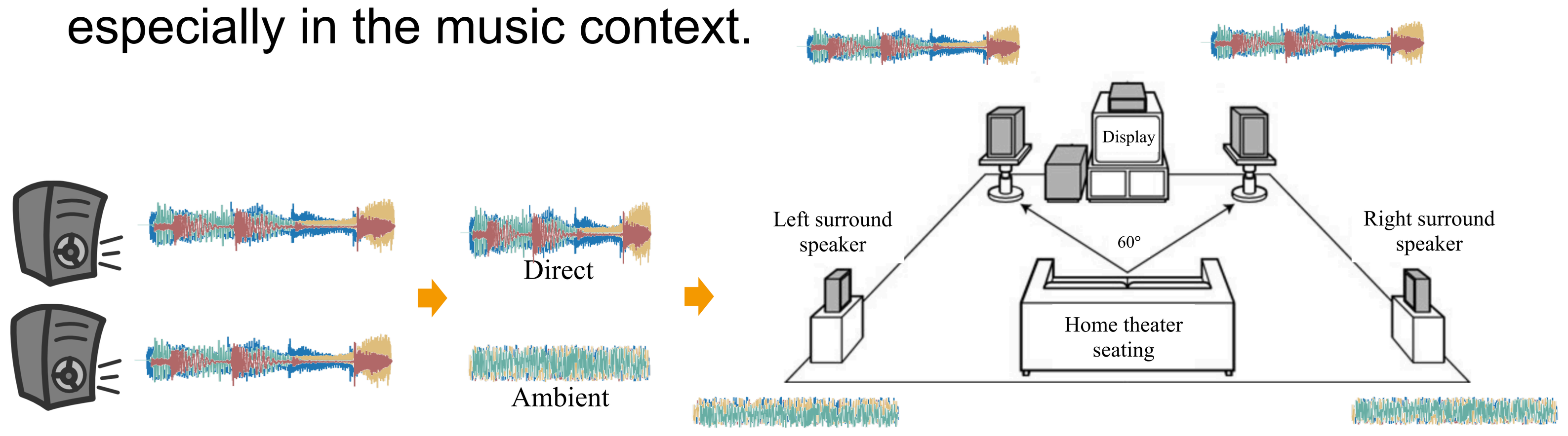
Upmixing via Style Transfer: A Variational Autoencoder for Disentangling Spatial Images and Musical Content

Haici Yang[1*], Sanna Wager[2], Spencer Russell[2], Mike Luo[2], Minje Kim[1,2*], Wontak Kim[2]
[1] Indiana University, Dept. of Intelliegent Systmes Engineering, Bloomington, IN, USA
[2] Amazon Lab126, Cambridge, MA and Sunnyvale, CA, USA * The work was done at Amazon

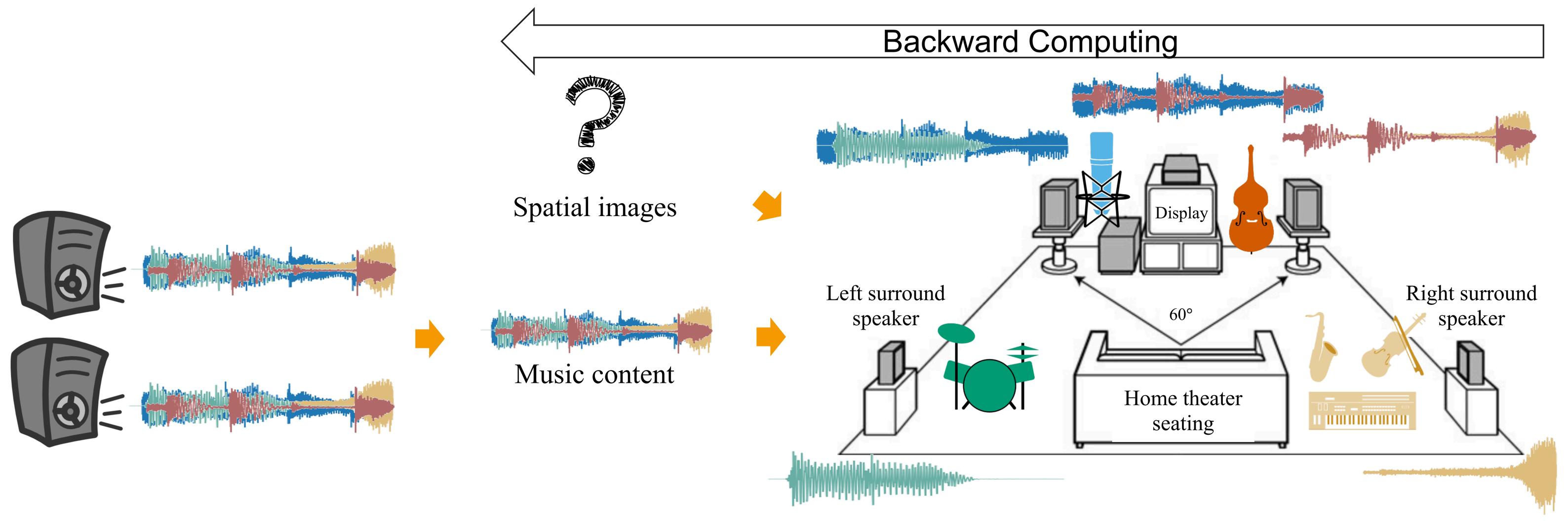
Paper ID 4030

Motivation

- Music upmixing: automatic conversion of stereo music to 5 channel surround material.
- **Conventional upmixing algorithms:**
  - Decompose the stereo into direct and ambient components
  - We believe they don't provide the optimal surrounding effect, especially in the music context.



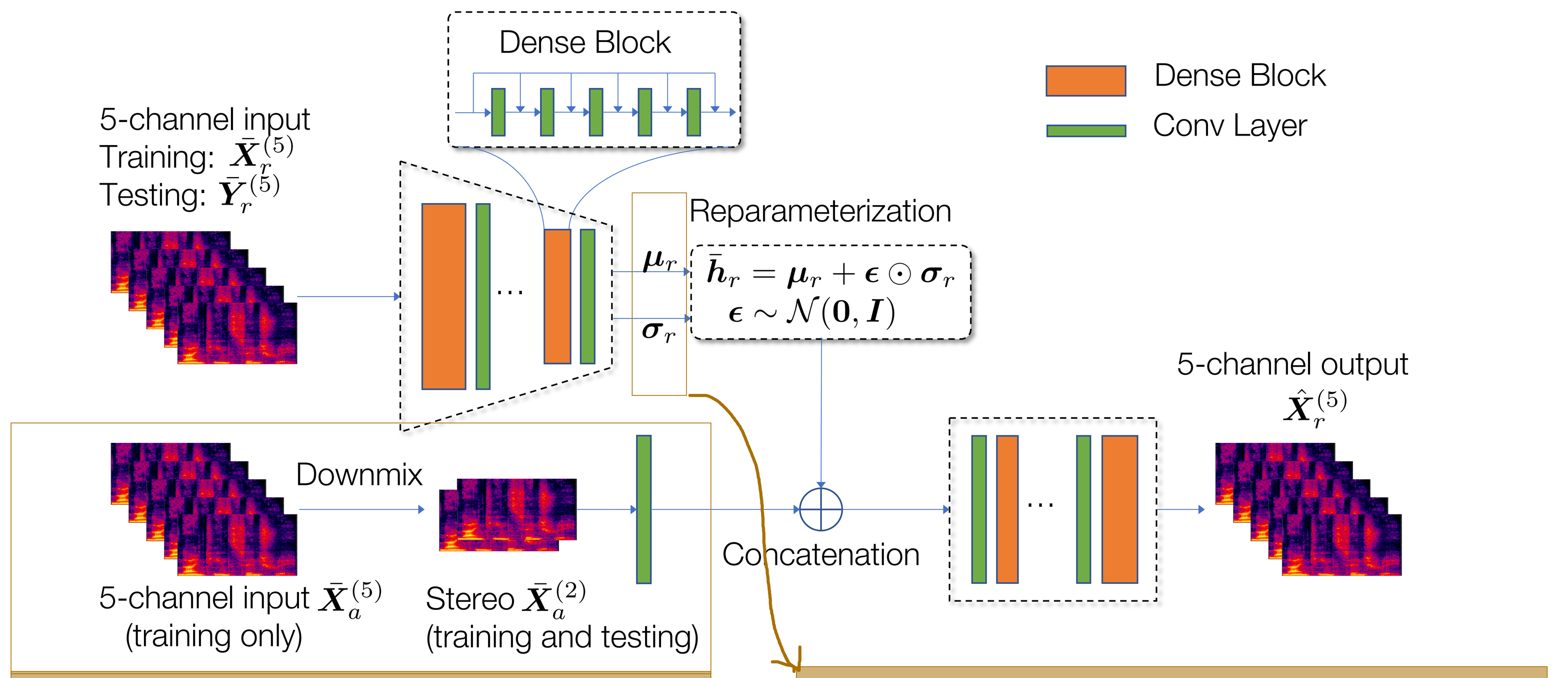
- **Our proposal:**
  - A virtual sound space for music playback scenario, where instruments are rendered at different spatial locations, perceptually.



- What is the correct way to place the instruments?
  - There is no "golden" answer to the question.
  - We could ask for users' guidance – too tedious.
  - Or reuse spatial information transferred over from existing 5ch
    - Entails that the spatial images and the music content are independent from each other.
- We compute this upmixing process backwards.
  - From a well established 5ch signal, we want to find a latent representation and disentangle these information in there.

Model

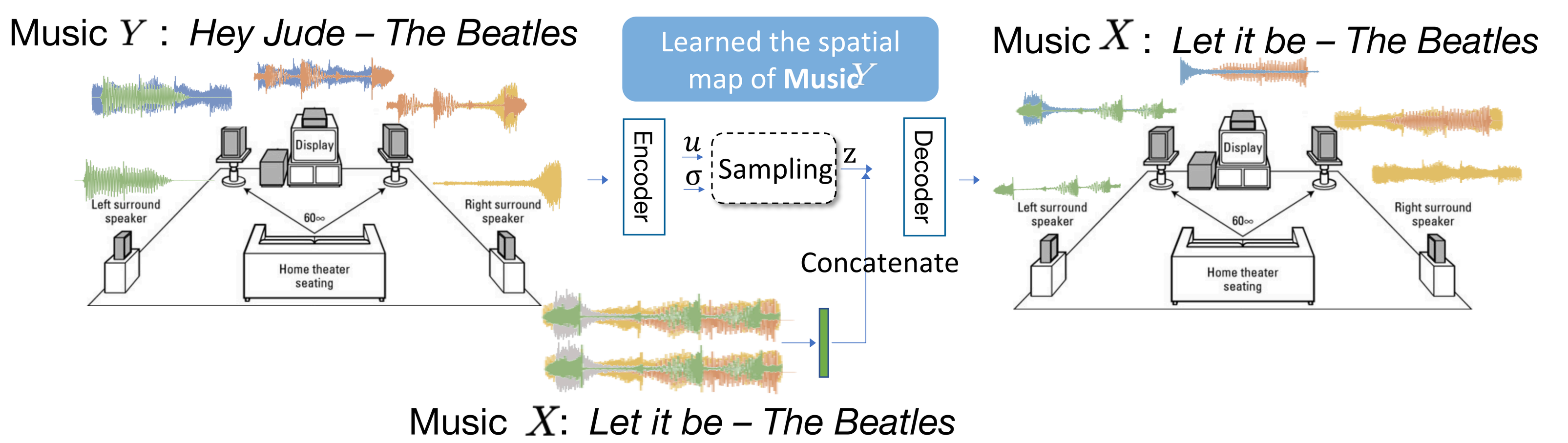
- **Model architecture**
  - Variational auto-encoder (VAE): Input and output are 5ch signal
  - We want the latent space to capture spatial information exclusively
  - We use a Densenet-like architecture for both encoder and decoder, to help the information flow during backpropagation.
- We make two main adaptions on the original VAE model:
  A. An extra stereo input into the decoder;
  B. We make the features in the latent space small enough, so that they can not capture any musical related content, but the spatial maps.



A. Now that music information can be completely provided by the stereo input, the only essential thing that bottle-neck should offer is the spatial information.

B. We assume that spatial maps are mostly constant, and the musical information is exponentially larger than the spatial information.

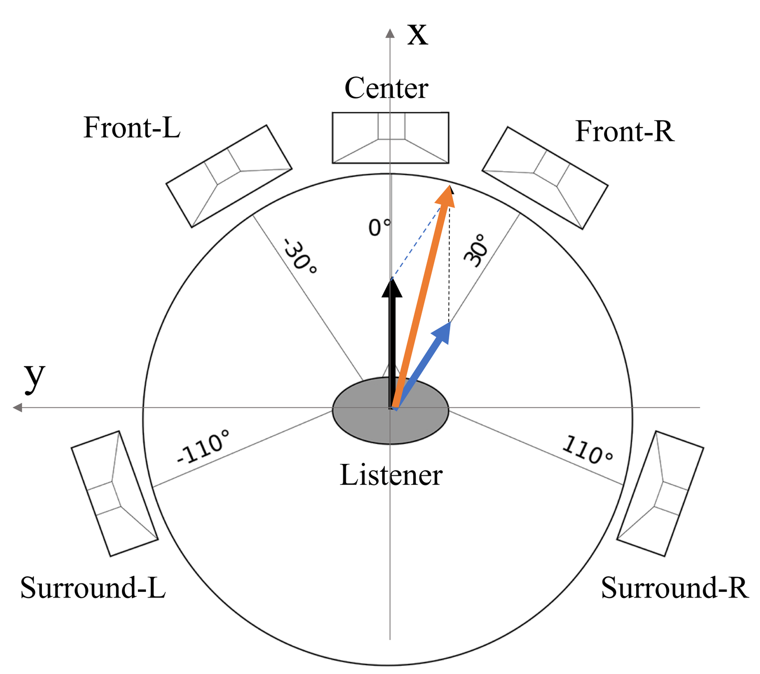
- **Model test**
  - **Style transfer-based upmixing** extracts the spatial images from music Y through the encoder to the latent variables. The variables are then fed into the decoder, together with the 2ch music X, to generate a 5ch music X that has Y's spatial map.



  - **Blind upmixing** uses random spatial images sampled from the latent space to generate the 5ch output.

Data Building

- We need a ground-truth 5-ch dataset, of which the instrument-specific spatial images are known and can be controlled. Current datasets cannot meet this requirement.
- Therefore, we build our own 5ch dataset using MUSDB18, by means of vector base amplitude panning (VBAP).
- We place the speakers per ITU's standards
- For each instrument, we first specify a virtual source direction, and then pan it independently using the two adjacent speakers towards the desired coming direction.



Experiment and Results

- **Visualization of the learnt latent space - Figure 1**
  - In Fig. 1a and Fig. 1b
    - Each line represents one single 50-dimension latent vector.
    - The latent variables highly corrected to the spatial maps and are invariants to the music content.
  - In Fig. 1c and Fig. 1d,
    - Each dot represents a dimension-reduced latent vector.
    - When colored by the panning method (Fig. 1d), the latent space is well structured.
    - Indicates that the encoder extracts music-invariant spatial features successfully.
- **The following evaluations compare the performances of style transfer-based upmixing and those of a baseline,**
  - We build the baseline by spreading each channel in the stereo to the front and rear channels of the same side in the 5-channel output.
- **Objective evaluation – Table 1**
  - SD-SDR: Scale dependent source to distortion ratio.
  - ΔAngle°: Difference between desired and output virtual angle for each source.
  - WILD: Wasserstain distance between the distribution of ground-truth inter-channel level differences and that of predicted ones.
- **Subjective evaluation – Figure 2**
  - In an ABX test, participants chose the one similar to the ground truth in terms of the incoming directions of the different sources and the overall spatial images.
  - The box plot shows the percentage of the votes which prefer style-transfer upmixing than the baseline.

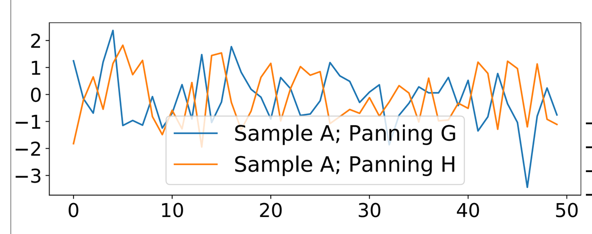


Figure 1a

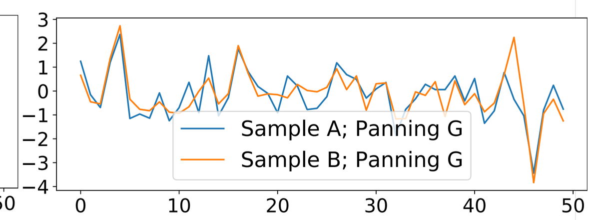


Figure 1b

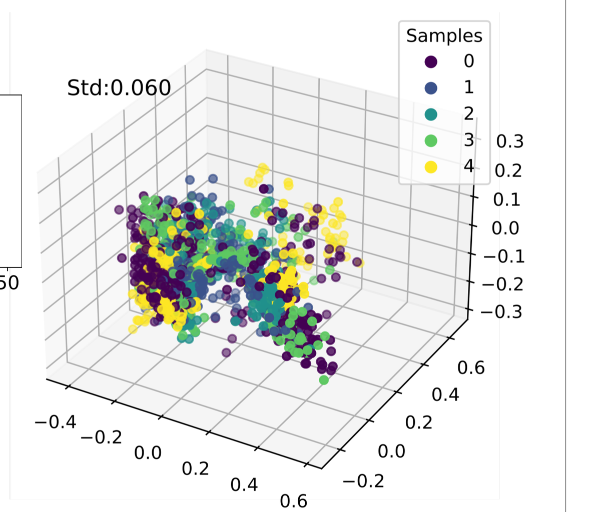


Figure 1c

	SD-SDR (dB)	Δ Angle (°)	WILD
Style-transfer	**8.71±1.50**	**19.5±15**	**58.02±13.78**
Baseline	4.20±0.71	49.75±19.75	70.84±10.03

Table 1 Our style-transfer upmixing outperforms the baseline over all criterions

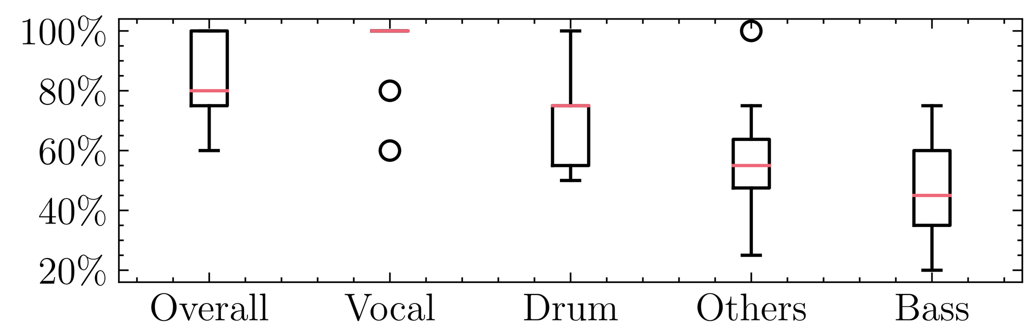


Figure 2 The vocal output from the style-transfer upmixing is best rated

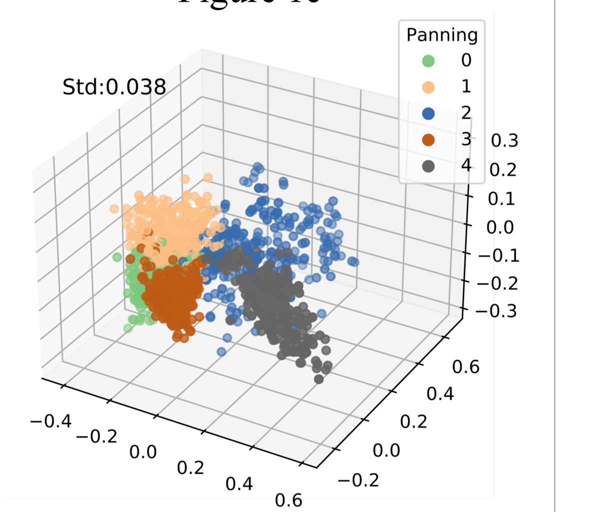


Figure 1d

Source codes and demo:
https://saige.sice.indiana.edu/research-projects/generative-upmixing/