Audio Query-based Music Source Separation

SEOUL NATIONAL UNIVERSITY

*Jie Hwan Lee, *Hyeong-Seok Choi and Kyogu Lee
(*: equal contribution)

MARG MUSIC & AUDIO RESEARCH GROUP

Problems (P) and Goals (G)

P: Most of the deep learning based music source separation models are dedicated to one specific class.

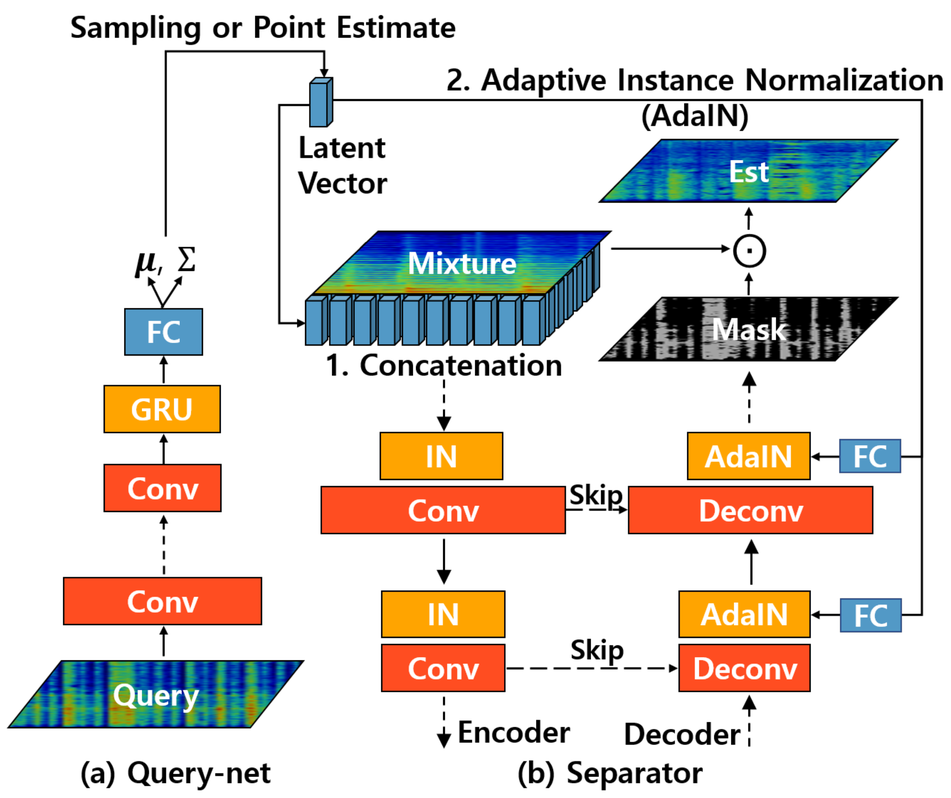
P: Most of the deep learning based models does not consider the inherent diverse characteristics of sources in a single class.

G: We propose a network for audio query-based music source separation that can directly encode the source information from a query signal.

G: Given a query and a mixture, the Query-net encodes the query into the latent space, and the Separator estimates masks conditioned on the latent vector, which is then applied to the mixture for separation.

Proposed method

Query-based Source Separation



- ✓ Query-net ($\mathbb{Q}$) is composed of 6 CNN and GRU layer.
- ✓ Separator($\mathbb{S}$) is a U-net based network.
- ✓ To effectively pass the summary of the query signal to Separator, we applied two methods.
  1. Concatenating the latent vector along the channel dimension of the input mixture spectrogram expecting the summarized information to be delivered from the start.
  2. Using AdaIN in the decoding stage of Separator

Training

- To design the proposed framework, we borrow the formulation of conditional variational autoencoder(cVAE). 1. Reconstruction loss, which is one of the objectives of cVAE, is used to force the output of $\mathbb{S}$ to be dependent on the encoded latent vector $\boldsymbol{z}$, 2. KL-divergence loss is used to make the distribution of $\boldsymbol{z}$ be close to the Gaussian distribution, 3. to enforce the output of Separator to be more dependent on the latent vector, we adopted latent regressor loss.

$$1.\ \mathcal{L}_R = \mathbb{E}_{S_T\sim p(S_T),M\sim p(M),\boldsymbol{z}\sim\mathbb{Q}(S_T)}[\|S_T - \mathbb{S}(M,\boldsymbol{z})\|]_1$$

$$2.\ \mathcal{L}_{KL} = \mathbb{E}_{S_T\sim p(S_T)}[\mathcal{D}_{KL}(\mathbb{Q}(S_T\|\mathcal{N}(0,I))]$$

$$3.\ \mathcal{L}_{latent} = \mathbb{E}_{M\sim p(M),\boldsymbol{z}\sim p(\boldsymbol{z})}\|\boldsymbol{z} - \mathbb{Q}(\mathbb{S}(M,\boldsymbol{z}))\|_1$$

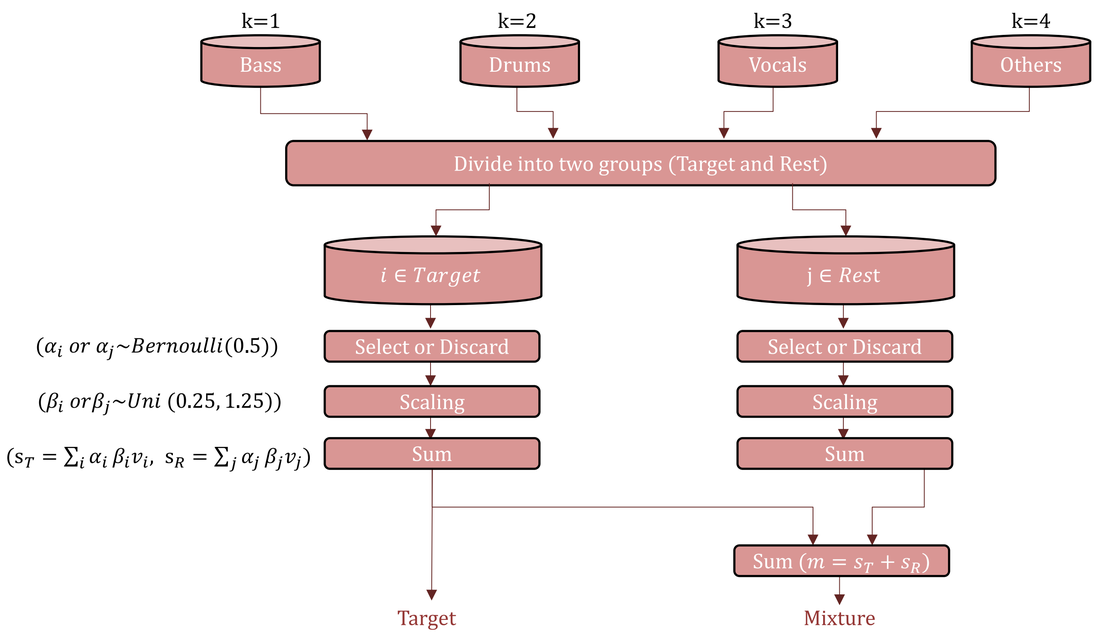
(S_T: target source, M: mixture)

Dataset & Sampling strategy

Dataset

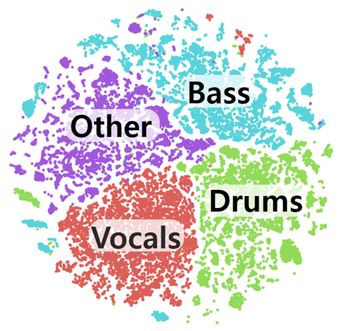
- MUSDB18 dataset which includes 4 instrument classes (bass, drums, vocals, others) was used for training and evaluation.

Sampling strategy



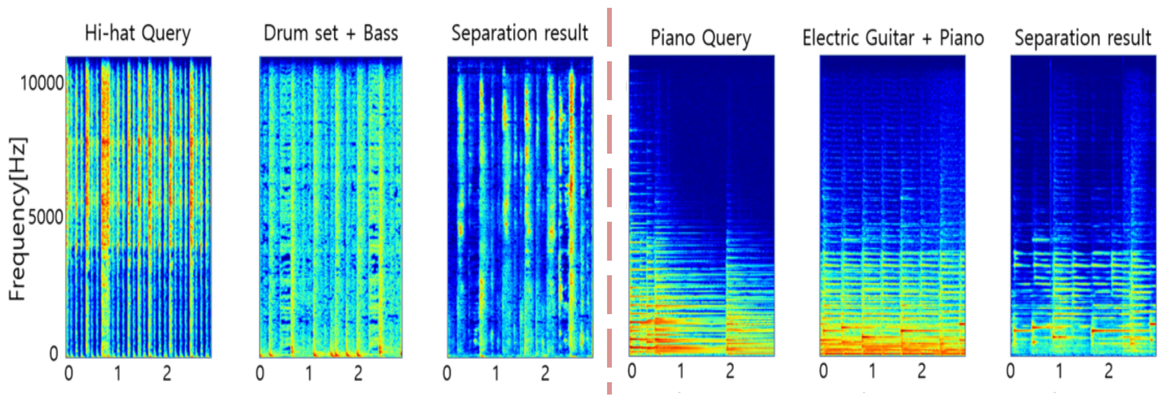
Experiments

1. t-sne plot of latent space



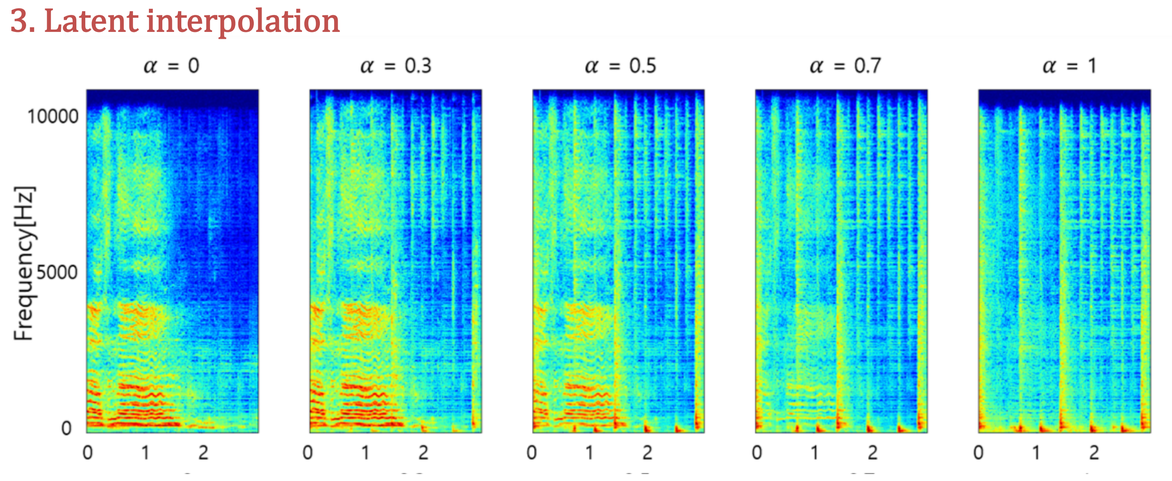
- The t-sne plot shows that each instrument classes with similar characteristics are grouped together, showing that the Query-net is able to provide the useful information to Separator

2. Manually targeting a specific sound source



- An audio query of hi-hat and piano were given to the mixtures of (hi-hat + kick drum + bass) and (piano + electric guitar).
- The noticeable fact is that we trained our method only with the MUSDB18 dataset, which has no hierarchical class label information besides the coarsely defined labels of sources such as `vocals', `drums', `bass' and `other'.
- Although our method was never trained to separate the subclass from the mixture, it was able to separate hi-hat and piano from the mixture, which can be referred to as a zero-shot separation.

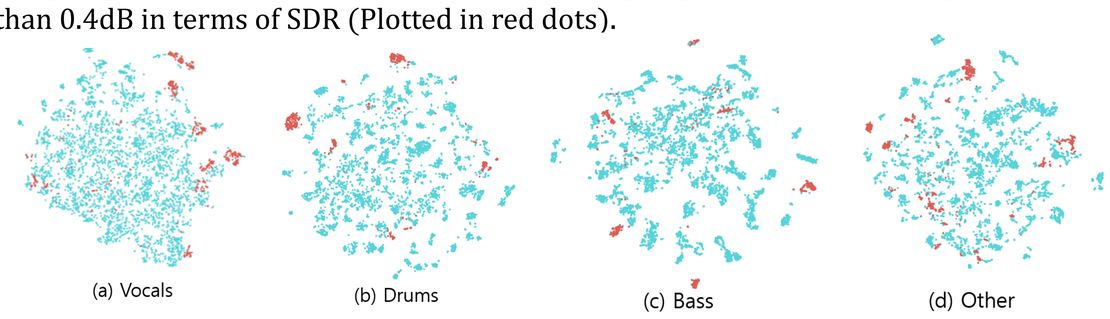
3. Latent interpolation



- To conduct a latent interpolation experiment, we computed the mean vector of each source.
- Above figure shows the interpolation results between vocals and drums, and α denotes the weight of slerp interpolation (e.g., $\frac{\sin(1-\alpha)\theta}{\sin\theta}\cdot\bar{z}_{vocal} + \frac{\sin\alpha\theta}{\sin\theta}\cdot\bar{z}_{drums}$).
- The intensity of separated instruments changes as the weight α changes. These experimental results show that our method can generate continuous outputs just by manipulating a latent space.

4. Iteratively separating source

- Iterative method is a technique that automates the query-based framework in an iterative way, which can be helpful under the harsh condition where the target sources are far from generic class.
- The Iterative method is done as follows,
  1. Separate the target source using the mean vector of certain sound class $\bar{z}$.
  2. Re-encode the separated source into a latent space expecting the re-encoded latent vector to be closer to the target latent vector.
  3. Separate the target source using the re-encoded latent vector.
- The results (Single step → Iterative) are as follows, `vocals': 4.84 → 4.90, `drums': 4.31 → 4.34, `bass': 3.11 → 3.09, and `other': 2.97 → 3.16.
- We grouped the tracks in the test set into two groups, the ones which gained more than 0.4dB in terms of SDR (Plotted in red dots).



5. Algorithm comparison

	Vocals	Drums	Bass	Other
STL2	3.25	4.22	3.21	2.25
WK	3.76	4.00	2.94	2.43
RGT1	3.85	3.44	2.70	2.63
JY3	5.74	4.66	3.67	3.40
UHL2	5.93	5.92	5.03	**4.19**
TAK1	**6.60**	**6.43**	**5.16**	4.15
Ours (mean)	4.90	4.34	3.09	3.16
Ours (GT)	5.48	4.59	3.45	3.26

Table. Median scores of MUSDB18 dataset

- (mean) denotes the results of using mean latent vector.
- (GT) denotes the results of using the ground truth spectrogram as a query which shows the upper bound of proposed method