Problems using deep generative models for probabilistic audio source separation

Searching for PhD in these topics!

Maurice Frank
maurice.frank@posteo.de

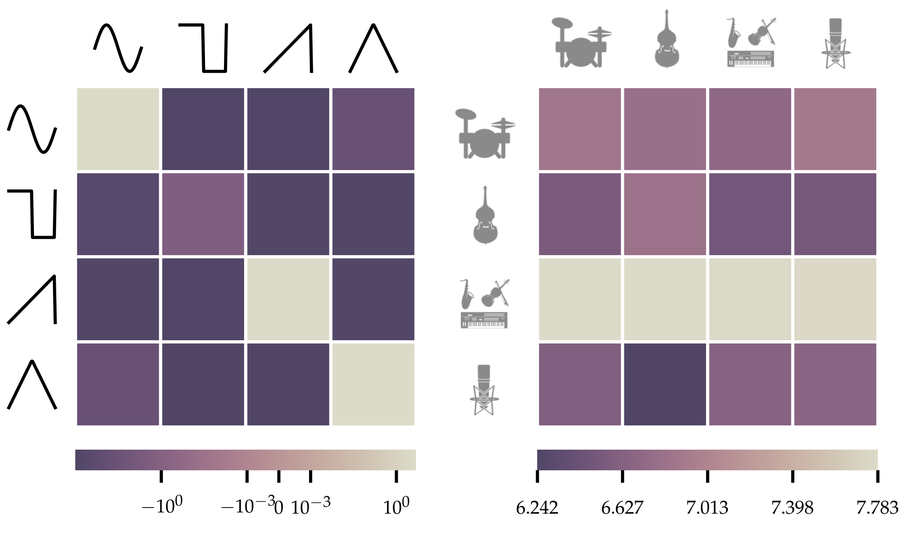
Maximilian Ilse
m.ilse@uva.nl

Idea: Train generative prior models for different sources of musical sounds. Use those probability density functions to solve multi-instrument tasks like audio source separation. For example we can pose source separation as the sampling procedure of finding the set of fitting samples from each distribution. This could be achieved with Langevin dynamics, but ...

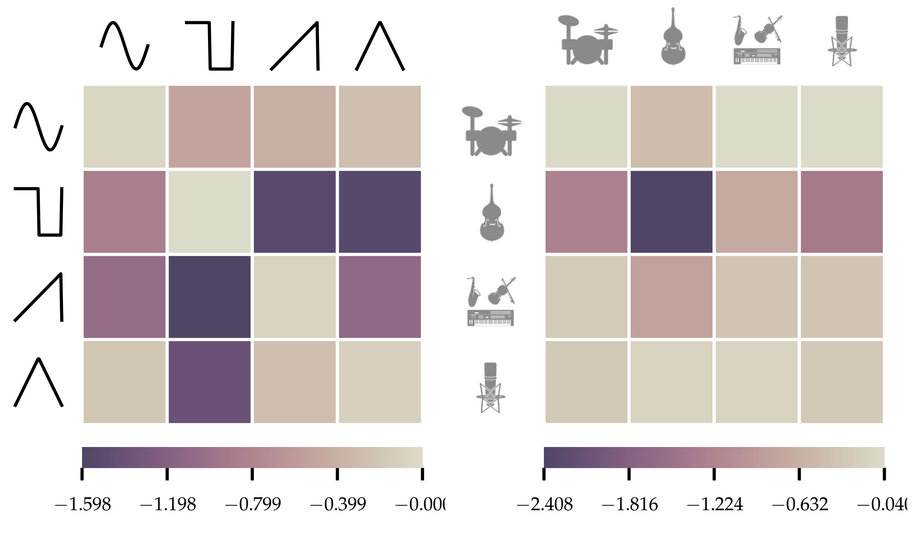
What happens if we train a generative model for sound sources separately?

Two types of model
each trained on real instruments and toy data

FloWaveNet (normalizing flow)



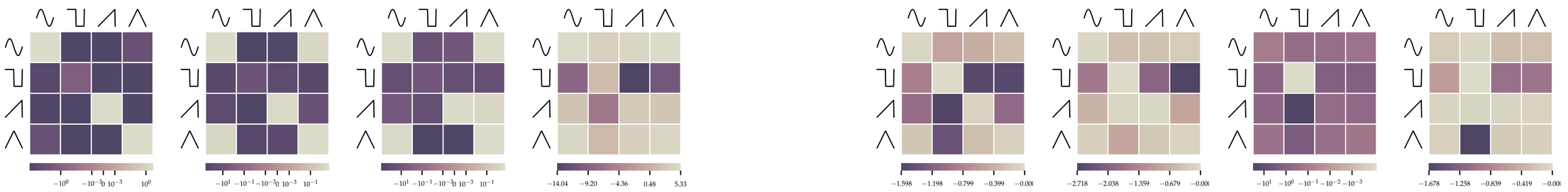
WaveNet (autoregressive)



The flow model can discriminate between in and out-of distribution samples for the simple toy-dataset while the wavenet can only slightly so.
Both confuse the out-of-distribution samples if trained on real world musical sounds.
Even more problematic models are too peaked around true samples to be used as prior distributions. We find the likelihood of in-class samples quickly reduces with added even the smallest amount of noise.

Recent work has suggested that the peakedness of the learned distribution can be allivated by training the network with noised input samples thereby approximating the corresponding smoothed out distribution.

Add noise to data $\overset{?}{\Rightarrow}$ Smooth out distribution



Notice that with even the smallest amount of added Gaussian noise the discriminative power in both models decreases rapidly.

Therefore we can conclude that, yes noising the input data does smooth the latent distribution. But it does level the distribution out to a point were a previously discriminative distribution is not discriminative anymore.

For your generative audio priors, pick one:

Discriminative *but* peaked

or

Smooth *but* not-discriminative