GAN for Audio Source Separation

Shaotian Chen
Department of ECE

Yueyi Yao
Department of ECE

Haikang Tan
Department of AME

ABSTRACT

Singing voice separation is a rising problem in the audio processing and machine learning area. The goal is to extract the voice track from the single original mixture audio.In this paper, we construct two GAN models to implement singing voice separation and compare the results with another two models: encoder-decoder with GAN and encoder-LSTMRNN-decoder with GAN

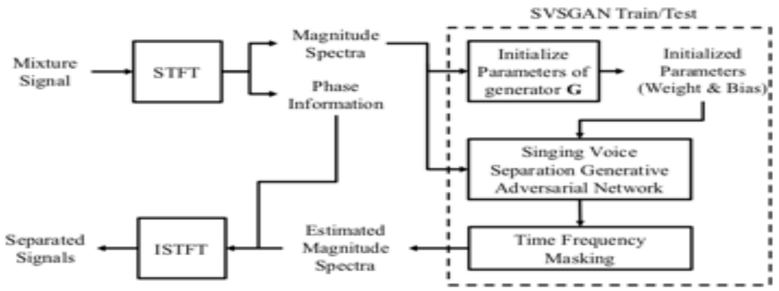

Figure 1 Singing voice separation based on GAN[2]

INTRODUCTION

Generative adversarial networks (GANs) are deep neural net architectures comprised of two nets, pitting one against the other (thus the “adversarial”). GANs were introduced in a paper by Ian Goodfellow and other researchers at the University of Montreal, including Yoshua Bengio, in 2014, GANs’ potential is huge, because they can learn to mimic any distribution of data. That is, GANs can be taught to create worlds eerily similar to our own in any domain: images, music, speech, prose. So we apply it in the audio area to separate the clean singing voice from the mixture. Here is the architecture of GAN when applied to audio.

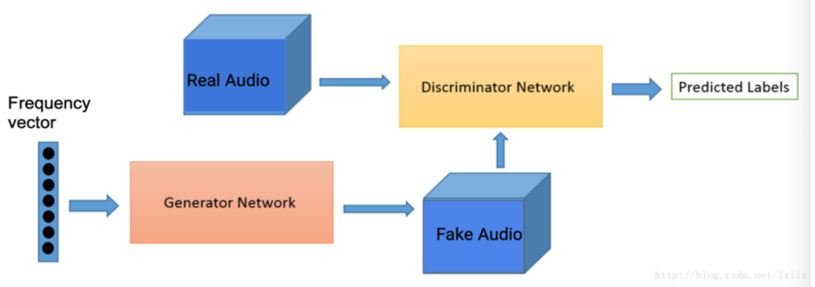

Figure 2 Architecture of GAN[1]

METHOD

In this method, we first take magnitude spectrogram via STFT (Short Time Fourier Transform) as input instead of image. our methods are based on [2] and [3] and we two pairs of architecture to compare with each other to see the performance. For evaluation, we use amounts of songs in DSD dataset to prove the effectiveness and to see the

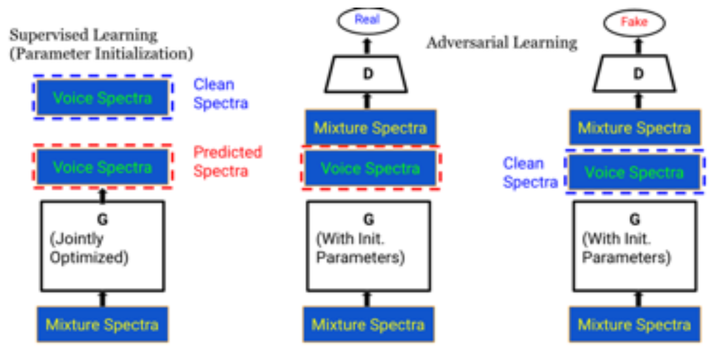

Figure 3 Architecture of this method[2]

- Network Selection:1.encoder-decoder. 2. encoder-lstm-decoder 3. encoder-decoder with GAN. 4. encoder-lstm-decoder with GAN
- Pre-processing: Hanning window of n samples (2048) with a n/2 hop size to segment it into T frames.
- Loss function for Generative Adversarial Network
  - Training objective function: Generalized KL divergence between the generated spectra and the clean spectra in supervised learning with Gaussian initialization.

$$D(\boldsymbol{X}\|\boldsymbol{Y}) = \sum_{ij} x_{ij}\left(\log\left(x_{ij}+\epsilon\right) - \log\left(y_{ij}+\epsilon\right)\right) - x_{ij} + y_{ij}$$

  - In adversarial learning, the whole loss function is below, where this model could learn data distribution from the clean spectra and the the mixture spectra.

$$\min_G \max_D V_{SVSGAN}(G,D) = E_{x\sim P_{data(z,s_c)}}[logD(s_c,z)] + E_{z\sim P_{G(z)}}[\log\left(1 - D(G(z),z)\right]$$

- Phase Reconstruction: Griffin-Lim algorithm

EXPERIMENT

- Dataset:
  - Dataset 1: DSD100: A dataset of 100 full lengths music tracks of different styles along with their isolated drums, bass, vocals and other stems. it contain two folder-train set, dev set, test set. we train on the train set, optimized in dev set, and then test in the test set.
- Result
  - Experiment 1: The result of SDR of encoder-decoder and encoder-decoder with GAN respectively
  - Experiment 2: The result of SDR of encoder-lstm-decoder and encoder-lstm-decoder with GAN respectively

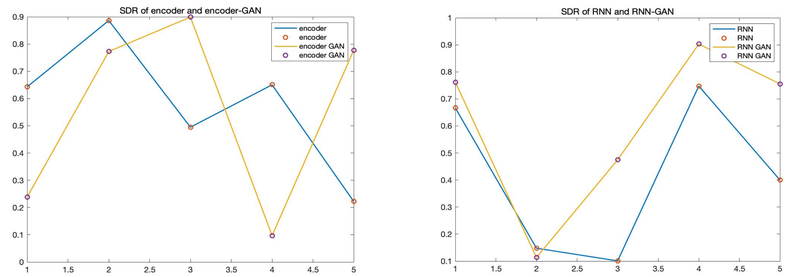

Figure 4 SDR of encoder-decoder and encoder-decoder with GAN(normalized) ; SDR of encoder-LSTM-decoder and encoder-LSTM-decoder with GAN(normalized)

DISCUSSION

The model is inspired by generative adversarial network, which is used in image generation.Since singing voice separation could also be regarded as a task of generation from the mixture given us. So we could apply GAN model in the singing voice separation. The goal of this model is to get a good generator that could deceive the discriminator. By adding the discriminator, we could improve our performance on the two generator we develop in the last semester.

REFERENCE

[1]Ian Goodfellow Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.
[2]Zhe Cheng Fan SVSGAN arXiv:1710.11428, 2017
[3]Y.Yan and Z.Duan HW5, 20118