SVSGAN: SINGING VOICE SEPARATION VIA GENERATIVE ADVERSARIAL NETWORK

Zhe-Cheng Fan, Yen-Lin Lai, and Jhy-Shing Roger Jang
{lambert.fan, andy.lai, jang}@mirlab.org
Dept. Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Singing Voice Separation (SVS)

- Goal: Extract singing voice from the polyphonic audio music
- Restriction: Only one channel for analysis
- Approach: Deep neural network (DNN)

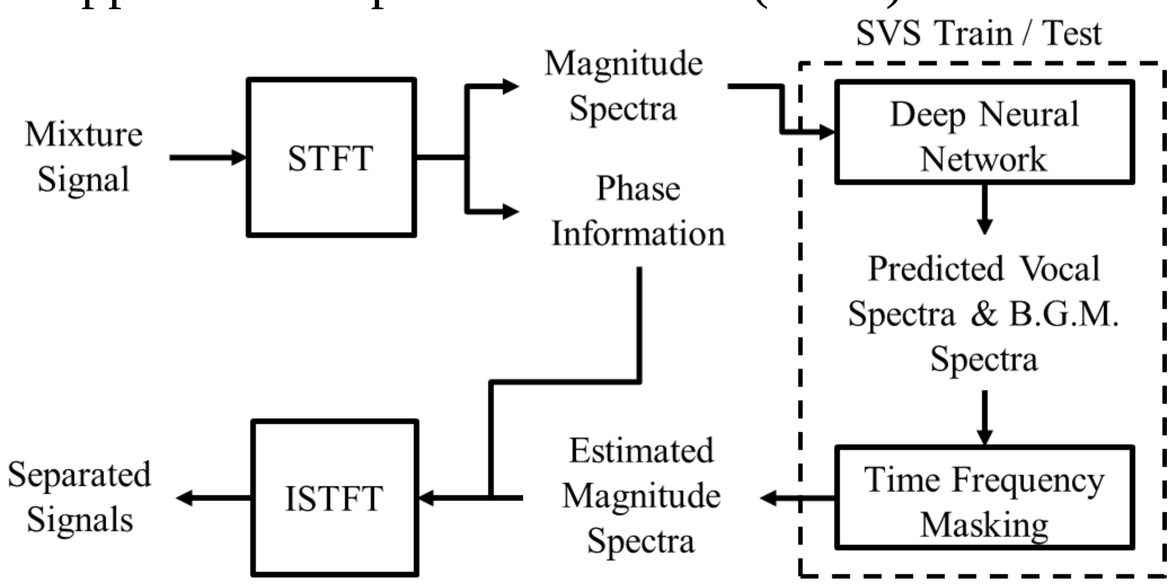

Fig. 1. Block diagram of SVS by using DNN

SVS via Generative Adversarial Network

- Parameters are initialized in a supervised setting
- Performance is optimized during adversarial learning
- Framework: Two conventional DNNs, **G** and **D**

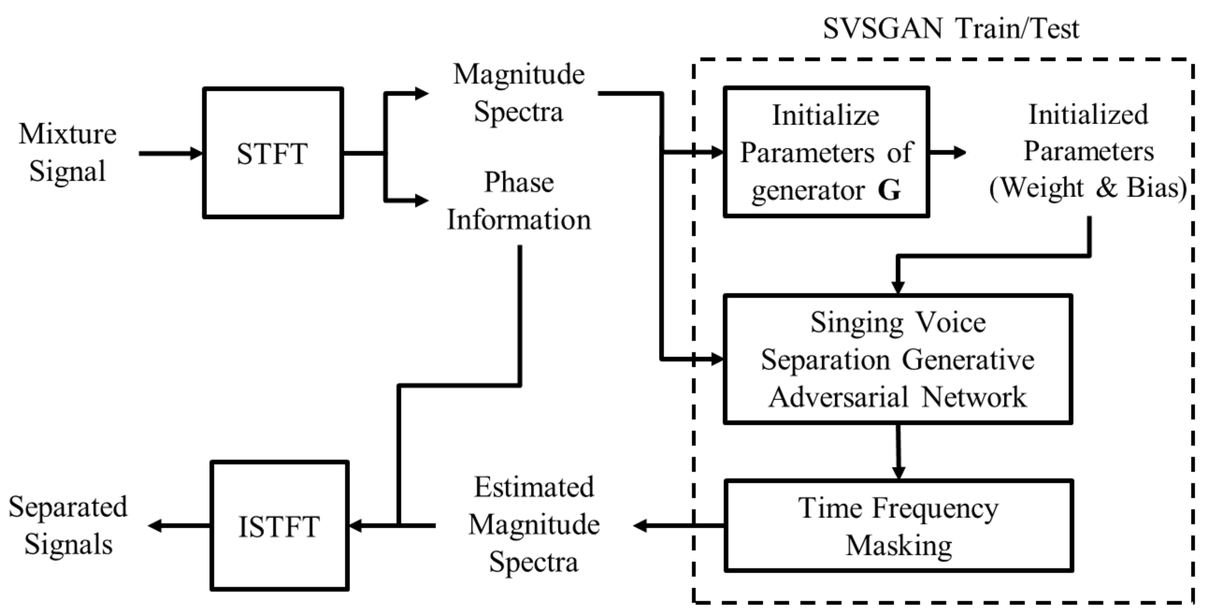

Fig. 2. Block diagram of SVSGAN

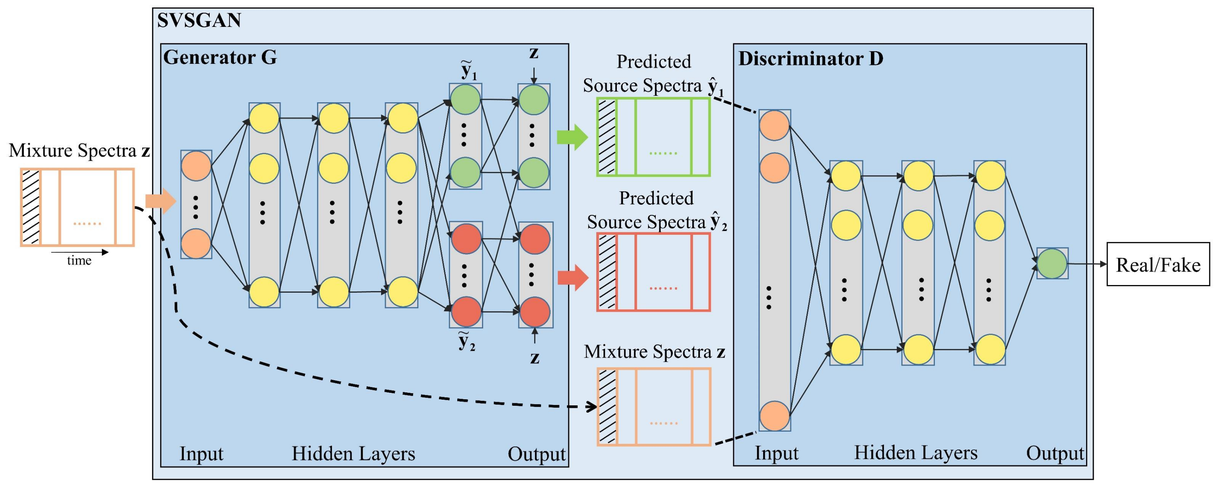

Fig. 3. The proposed SVSGAN framework. Each spectrum is considered to be a sample vector coming from a distribution of spectra.

Experimental results

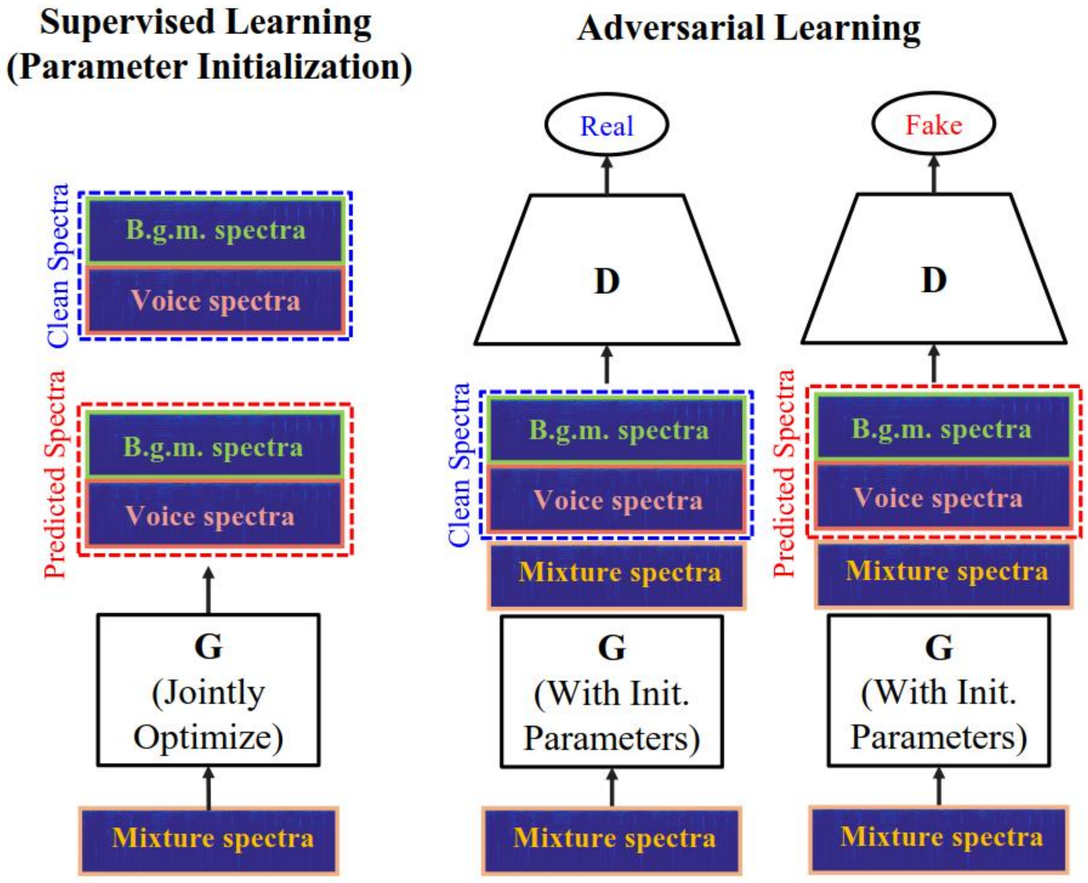

Fig. 4. SVSGAN training process

MIR-1K Dataset			
Model	SDR	SAR	SIR
DNN (baseline)	6.57	10.14	9.84
SVSGAN (V+B)	6.69	10.32	9.86
SVSGAN (V+M)	6.73	10.28	9.96
SVSGAN (V+B+M)	6.78	10.29	10.07
IBM (upper bound)	13.92	14.80	21.96
iKala Dataset			
Model	SDR	SAR	SIR
DNN (baseline)	9.74	11.72	14.99
SVSGAN (V+B)	10.15	12.48	14.72
SVSGAN (V+M)	10.22	12.78	14.41
SVSGAN (V+B+M)	10.32	12.87	14.54
IBM (upper bound)	12.30	14.10	23.70

Table 1. Vocal results (in dB) of conventional DNN and SVSGANs

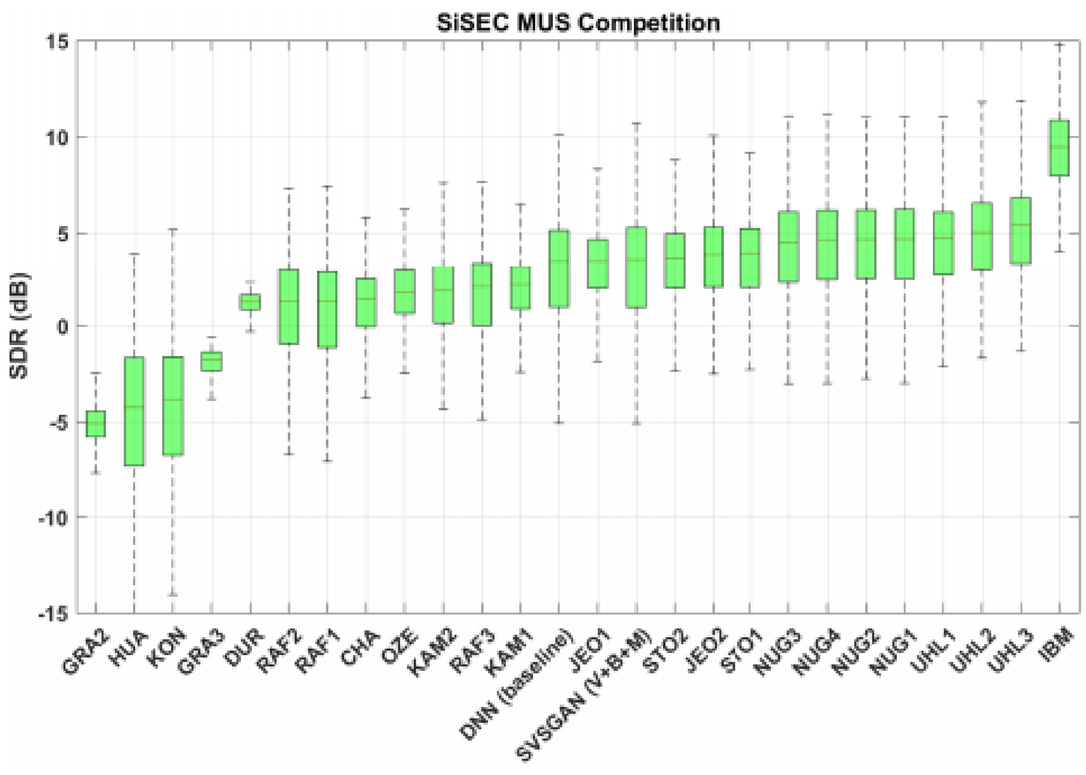

Fig. 5. Vocal results on the Test part of the DSD100 dataset