신경망 기반 음원 분리 시스템의 학습 속도 향상을 위한 음역대 강조 기법

김민석*, 최우성*, 정순영*
*고려대학교 컴퓨터학과
rlaalstjr47@korea.ac.kr, ws_choi@korea.ac.kr, jsy@korea.ac.kr

Frequency Range Enhancement for Faster Convergence of Neural Music Source Separation Systems

Min-Seok Kim*, Woo-Sung Choi*, Soon-Young Jung*
*Dept. of Computer Science, Korea University

요 약

여러 악기가 섞여 있는 음원으로부터 원하는 악기 소리를 추출하는 음원 분리 기법 중 최근 신경망 기반 시스템이 활발히 연구되고 있다. 악기마다 고유의 음역대를 가진다는 사실에 감안하여, 연구진은 기존 음원 분리 신경망에 적은 수의 학습 파라미터를 추가하여 학습 속도를 대폭 향상시킬 수 있는 음역대 강조 기법을 제안한다.

1. 서론

음원 분리(Musical Source Separation)란, 여러 악기 소리로 이루어진 혼합 음원으로부터 원하는 악기 소리만으로 이루어진 단일 악기 음원을 추출하는 작업이다. 대중음악으로부터 목소리를 분리해내거나 반주만 남기는 작업이 대표적이다.

최근에 들어 딥러닝을 통한 음원 분리가 활발히 연구되고 있다 [1,2,3,4]. 이들은 대부분 시간에 따른 각 주파수 세기의 변화를 표현해주는 spectrogram을 학습 데이터로 사용한다. 이러한 기존 음원 분리 신경망의 더욱 효율적인 학습을 위해, spectrogram에 포함된 불필요한 주파수를 제거하고 음원 분리에 중요한 역할을 하는 주파수일수록 세기를 증가시켜 주는 "음역대 강조" 기법을 본 논문에서 제안한다.

2. 제안 기법

2.1 기존 연구 및 기초 지식

신경망 기반 음원 분리 시스템은 supervised 방식이 대다수이며, 학습 데이터 instance는 혼합 음원과 그로부터 추출하고자 하는 단일 악기 음원의 쌍으로 이루어진다. 신경망은 임의의 혼합 음원을 받았을 때 그에 대응되는 단일 악기 음원을 예측하여 생성하도록 학습된다.

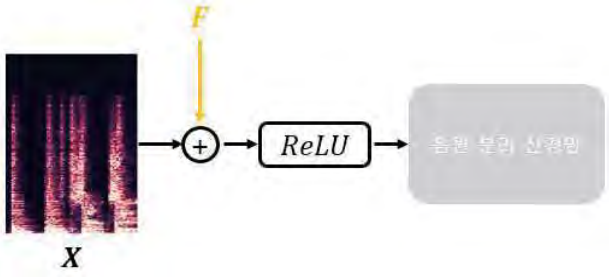
여기서 음원을 표현하는 방식에 따라 음원 분리 시스템을 크게 두 부류로 나눌 수 있다. 1차원 배열의 형태를 가지는 음악 신호(waveform)를 그대로 학습에 사용하는 방법([3,4]), 그리고 퓨리에 변환의 일종인 Short Time Fourier Transform(STFT)을 전처리 방식으로 이용하여 주파수 차원과 시간 차원으로 이루어진 2차원 배열(spectrogram)로 학습을 하는 방법([1,2])이 있다. 특히 spectrogram은 이미지와 유사한 특성을 가지기 때문에 이미지 처리 분야에서 개발된 CNN을 그대로 활용하여 준수한 성능을 얻을 수 있었다. 그림 1은 혼합 음원으로부터 목소리를 추출하는 spectrogram 방식의 신경망 기반 음원 분리의 예시이다.

(그림 1) Spectrogram 방식의 신경망 기반 음원 분리.

2.2 음역대 강조 기법

제안 기법은 spectrogram 방식으로 개발된 기존 신경망의 학습을 개선시키는 방법이다. 혼합 음원의 spectrogram이 음원 분리 망을 통과하기 전에, 음원 분리에 필요한 주파수만 남기도록 하는 것이다. 어떤 주파수를 강조하고 어떤 주파수를 거를지는 추가 파라미터를 통해 학습되며, 이 학습은 기존 음원 분리 신경망의 학습과 동시에 end-to-end 방식으로 진행된다.

물론, 이러한 주파수 필터링은 사람의 수작업으로도 충분히 가능하다(downsampling, equalizer 활용 등). 하지만 음원 분리 신경망이 대상 악기의 음역대 안의 정보만 활용한다는 보장이 없기 때문에, 단순히 악기의 음역대가 아닌 음원 분리에 필요한 음역대를 학습할 수 있는 제안 기법이 이러한 전처리 방식과 차별화될 수 있다.

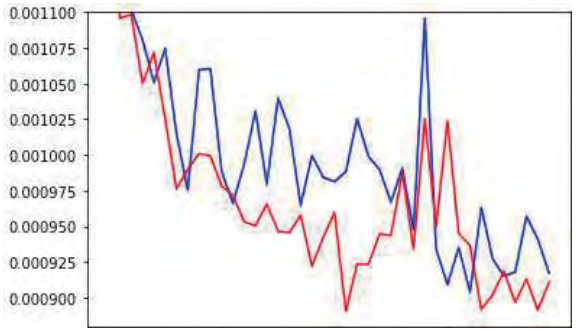


(그림 2) 음역대 강조 기법.

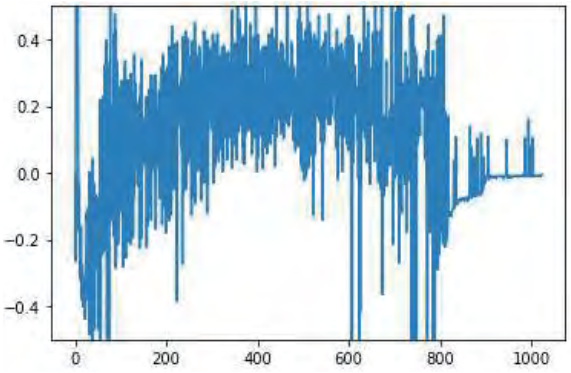
제안 기법은 그림 2처럼 표현될 수 있다. 신경망 input으로 들어온 mixture spectrogram $X \in R^{d_t \times d_f}$에 음역대 강조를 위한 추가 파라미터셋 $F \in R^{1 \times d_f}$을 broadcast하여 더해준 후, ReLU activation을 통과시켜 음수를 0으로 만들어준다. 단위 시간마다 같은 1차원 벡터 F를 더해줌으로써 시간 축과는 무관한 주파수 축만의 특징을 학습하도록 하였다. 결국 F는 절댓값이 큰 음수들을 활용하여 불필요한 주파수들을 제거하고, 중요한 주파수일수록 큰 양수를 더하여 강조하도록 학습된다(실험 평가 그림 4).

3. 실험 평가

목소리 음원 분리로 실험을 진행하였으며, 이를 위해 MUSDB[5] 데이터셋을 사용하였다. 수렴 속도 비교 실험에 사용된 기존 음원 분리 기법은 MDenseNet[1]이며 frame 수(d_t)를 제외한 모든 STFT 파라미터는 [1]에 기재된 것과 동일하다 ($d_t = 64$, $d_f = 1025$). 학습 loss function 및 validation metric은 모두 MSE를 사용하였다.



(그림 3) Validation loss의 수렴 속도 비교.



(그림 4) F 벡터 가시화.

그림 3은 음역대 강조로 인한 수렴 속도 향상을 보여준다. 청색 그래프는 MDenseNet, 적색 그래프는 이에 음역대 강조 단계를 추가한 망의 validation loss를 나타낸다. 각각 대략 7 epoch 동안 학습이 진행됐으며, 제안 기법의 경우 기존 기법과 비교했을 때 약 2 epoch 일찍 수렴한 것을 알 수 있다. 최종 수렴 loss는 심지어 더 낮은 경우도 많았으나, 분리된 목소리 샘플을 비교하여 감상했을 때 유의미한 차이가 느껴질 만큼은 아니었다.

마지막으로, 그림 4의 뒤집어진 포물선 형태를 통해 F의 파라미터셋이 지나치게 낮거나 높은 주파수를 걸러내는 방향으로 학습됐음을 알 수 있었다.

4. 결론

본 연구는 기존 음원 분리 신경망에 소수의 파라미터를 추가하는 것으로 학습 속도를 향상시킬 수 있음을 보여주었다. 또, 그림 4와 같은 가시화만으로도 사람이 학습 결과를 어느 정도 파악할 수 있다는 장점을 가진다. 더욱 일반화될 수 있는 기법인지를 확인하기 위해, 향후 다른 악기 혹은 다른 망을 대상으로도 실험을 진행할 예정이다.

5. 사사

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1F1A1062719).

참고문헌

[1] Naoya Takahashi, et al. “Multi-scale multi-band DenseNets for audio source separation", WAS-PAA, New York, 2017, pp. 261-265.

[2] Woosung Choi, et al. "Investigating Deep Neural Transformations for Spectrogram-based Musical Source Separation." arXiv preprint arXiv:1912.02591, 2019.

[3] Daniel Stoller, et al. “Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation", ISMIR, Paris, 2018, pp. 334-340.

[4] Alexandre Défossez, et al. "Music Source Separation in the Waveform Domain." arXiv preprint arXiv:1911.13254, 2019.

[5] Zafar Rafii, et al. “MUSDB18 - a corpus for music separation", 2017.